文章编号：1008-1542(2024)02-0150-09

基于监督学习的 HSK 阅读文本自动分级模型研究

任　梦[1]，王方伟[2]

(1. 河北师范大学文学院，河北石家庄　050024；
2. 河北师范大学计算机与网络空间安全学院，河北石家庄　050024)

摘　要：针对 HSK(汉语水平考试)各类阅读材料难度判定与等级对应中缺乏有效参照标准和分析工具的问题，以历年 HSK 真题阅读文本为研究对象，提取文本可读性特征，采用支持向量机、随机森林、极端梯度增强等 9 种监督学习算法，建立可将自选文本自动归类于相应 HSK 等级的模型，采用准确率、AUC 等多项指标评价各模型的分级效果，并选择最佳模型制成在线工具。结果表明，监督学习在 HSK 阅读材料文本分析及分级方面具有较高性能，9 种模型中极端梯度增强的分级效果最好，准确率为 0.913，AUC 为 0.994。建立的分级模型和在线工具能够以较高的准确率对 HSK 自选文本进行分级，帮助用户有针对性地遴选文本，提高学习效率。

关键词：自然语言处理；监督学习；HSK 阅读文本；可读性特征；分级模型

中图分类号：TP391.77　　**文献标识码**：A　　**DOI**：10.7535/hbkd.2024yx02005

Research on automatic grading model of HSK reading texts based on supervised learning

REN Meng[1]，WANG Fangwei[2]

(1. College of Chinese and Literature，Hebei Normal University，Shijiazhuang，Hebei 050024，China；
2. College of Computer and Cyber Security，Hebei Normal University，Shijiazhuang，Hebei 050024，China)

Abstract：Aiming at the problem that there are few effective reference standards and analysis tools available in classifying and grading Hanyu Shuiping Kaoshi (HSK) reading materials，with HSK reading texts in the past years as study object，the text readability features were extracted，and nine supervised learning algorithms，such as support vector machine，decision tree and extreme gradient enhancement，etc.，were employed to build a model that could automatically classify self-selected text to the corresponding HSK level. Multiple indicators such as accuracy and AUC were adopted to evaluate the grading effect of each model，and the best model was chosen to design an online tool. The results show that supervised learning has high performance in analyzing and grading HSK reading materials. Among the nine supervised learning models，extreme gradient enhancement is the best，with an accuracy of 0.913 and an AUC of 0.994. The grading model and online tool can grade HSK self-selected texts with high accuracy，help users select texts pertinently and improve learning efficiency.

收稿日期：2023-12-19；修回日期：2024-03-06；责任编辑：张士莹
基金项目：国家自然科学基金(61572170)；河北师范大学 2023 年度人文社会科学校内科研基金(S23AI001)
第一作者简介：任梦(1990—)，女，河北石家庄人，讲师，博士研究生，主要从事自然语言处理等方面的研究。
E-mail：olivia24rm@126.com

任梦，王方伟. 基于监督学习的 HSK 阅读文本自动分级模型研究[J]. 河北科技大学学报，2024，45(2)：150-158.
REN Meng，WANG Fangwei. Research on automatic grading model of HSK reading texts based on supervised learning[J]. Journal of Hebei University of Science and Technology，2024，45(2)：150-158.

Keywords：natural language processing；supervised learning；HSK reading text；readability feature；grading model

HSK(汉语水平考试)是一项国际标准化考试[1]。自 2022 年 11 月起，HSK 在 1—6 级基础上新增了 7—9 级考试，从不同层面考查应试者的综合能力。当前互联网文本信息规模庞大，内容丰富，用户可以非常容易地获取各类汉语阅读材料。但如何判断这些材料的难度，是否能与 HSK 等级相对应，往往靠的是个人经验，缺乏有效的参照标准和分析工具。监督学习属于机器学习的一种，指的是利用一组带有标签的数据，学习从输入到输出的映射，然后将这种映射关系应用到未知数据上，达到分类或回归的目的。目前已经有研究者将汉语可读性特征和监督学习应用到 HSK 阅读文本的分析中。江新等[2]以 HSK(5 级、6 级)阅读文本为实验材料，建立了包含相异词比率和虚词数在内的可读性公式，依据该公式计算得出的可读性分数与专家对文本难度的评定分数高度相关；杜月明等[3]基于文本可读性特征集合，引入特征选择算法，通过对比 6 种监督学习模型的效果，实现了 HSK 阅读文本可读性的自动评估，其结果表明支持向量机模型在评估中的表现最好。但是通过梳理发现这些研究存在以下问题：第一，研究内容主要是从特征选择、文本分析、优化算法等理论层面进行的，建立的公式和模型虽达到了较好的分析效果，但未能将其转化为学习者可以利用的工具[4]；第二，已有研究主要聚焦于分析教材和考试文本，未能详述如何将研究成果具体应用在课外或者自选阅读材料上[5]；第三，研究大多采用传统回归算法，部分使用监督学习算法的研究主要采用的是经典的支持向量机、朴素贝叶斯等算法，或是随机森林等 Bagging(又称袋装法)算法，文本分析结果的准确率有待进一步提高[6-7]。近年来在监督学习领域，极端梯度增强、梯度提升决策树等 Boosting(又称提升法)算法以更好的分类、泛化性能和更高的运行效率得到广泛应用[8-16]，但尚未应用于与 HSK 相关的分析中。

针对以上情况，本研究以历年 HSK 真题阅读文本为研究对象，利用包括 Boosting 在内的 9 种监督学习算法，筛选与 HSK 等级相关的可读性特征，建立可将自选文本自动归类于相应 HSK 等级的模型，帮助用户有针对性地选取文本材料。

1　研究内容

1.1　研究对象

本研究收集了 2010—2018 年出版的《汉语水平考试 HSK 真题集》，同时结合网络资源，收集真题 193 套。经扫描录入、光学字符识别和排版整理，共采集文本 560 520 字。由于 HSK 各等级题型不同，部分题目字数过少，部分选项为单个或者并列词汇，可读性特征不全，不利于文本分析，故本研究根据题目类型和字数，将每 5 题或 6 题合并为一条文本，并剔除部分选项，最终纳入文本 1 350 条，共 448 173 字。详细数据信息见表 1。

表 1　HSK 真题阅读文本数据信息

Tab. 1　HSK reading texts data information

HSK 级别	试卷数	总字数	剔除字数	纳入字数	文本条数	单文本平均字数
1 级	34	6 451	0	6 451	102	63
2 级	29	17 003	2 312	14 691	135	109
3 级	32	40 769	7 190	33 579	192	175
4 级	32	78 688	17 356	61 332	256	240
5 级	34	188 992	48 541	140 451	313	449
6 级	32	228 617	36 948	191 669	352	545
总计	193	560 520	112 347	448 173	1 350	332

1.2　研究方法

1.2.1　监督学习算法

本研究使用 9 种不同的监督学习算法对数据集进行分析，包括支持向量机(support vector machine，SVM)、决策树(decision tree，DT)、K 近邻(K-nearest neighbor，KNN)、随机森林(random forest，RF)、极端随机树(extra trees classifier，ETC)、梯度提升决策树(gradient boosting decision tree，GBDT)、轻量级梯度提升(light gradient boosting machine，LGBM)、自适应增强(adaptive boosting，AdaBoost)和极端梯度增强

(extreme gradient boosting,XGBoost)。使用 Python 3.11 软件 Sklearn 模块编写监督学习算法代码,依据 Z-Score 将各项数据进行标准化处理,便于对不同单位或量级的指标进行比较和加权。为尽量减少过拟合现象,在算法允许的情况下进行 5 倍交叉验证。同时,在同等数据条件下使用 SPSS 27.0 软件进行 Logistic 回归分析,比较监督学习和传统回归算法的差异。

1.2.2 评估方法

利用网格搜索法编写调参代码辅助调整模型参数,采用 5 倍交叉验证计算,使各模型均达到自身最优效果,计算各模型的准确率、精确率、召回率、平衡 F 分数(F1-Score)。为便于和以往研究相比较,本研究主要采用准确率评价模型分级效果。同时,由于受试者工作特征(receiver operating characteristic,ROC)曲线和其曲线下面积(area under curve,AUC)兼顾分类的正例和负例,可全面反映灵敏度和特异性的关系,是一个较均衡的评估指标,故本研究结合 AUC 值评估模型性能。本数据集各组间文本量差距较大,准确率、精确率、召回率和 F1-Score 采用结合样本权重的加权平均值,AUC 采用结合不同类别贡献大小的微平均值。以上过程均重复进行 5 次,取均值作为最后结果,以减弱随机抽样的偶然性,使结果更加稳定。

1.2.3 在线运行环境

选择分级效果最好的模型,对各可读性特征的重要度进行计算和排名。使用递归式特征消除(recursive feature elimination,RFE)计算最佳特征个数,结合特征重要度排名,选出与 HSK 等级最相关的特征。最后,使用效果最好的模型和最相关的特征重新进行训练,达到最佳分级效果。利用 Python 中的 Flask 框架编写 Web 代码,上传至云服务器,使分级模型可通过网页工具的形式使用。

本研究的具体流程见图 1。需要说明的是,由于 HSK 7—9 级考试开始较晚,尚未有官方出版发行的真题供参考研究,故本研究暂未纳入该级别的阅读文本。

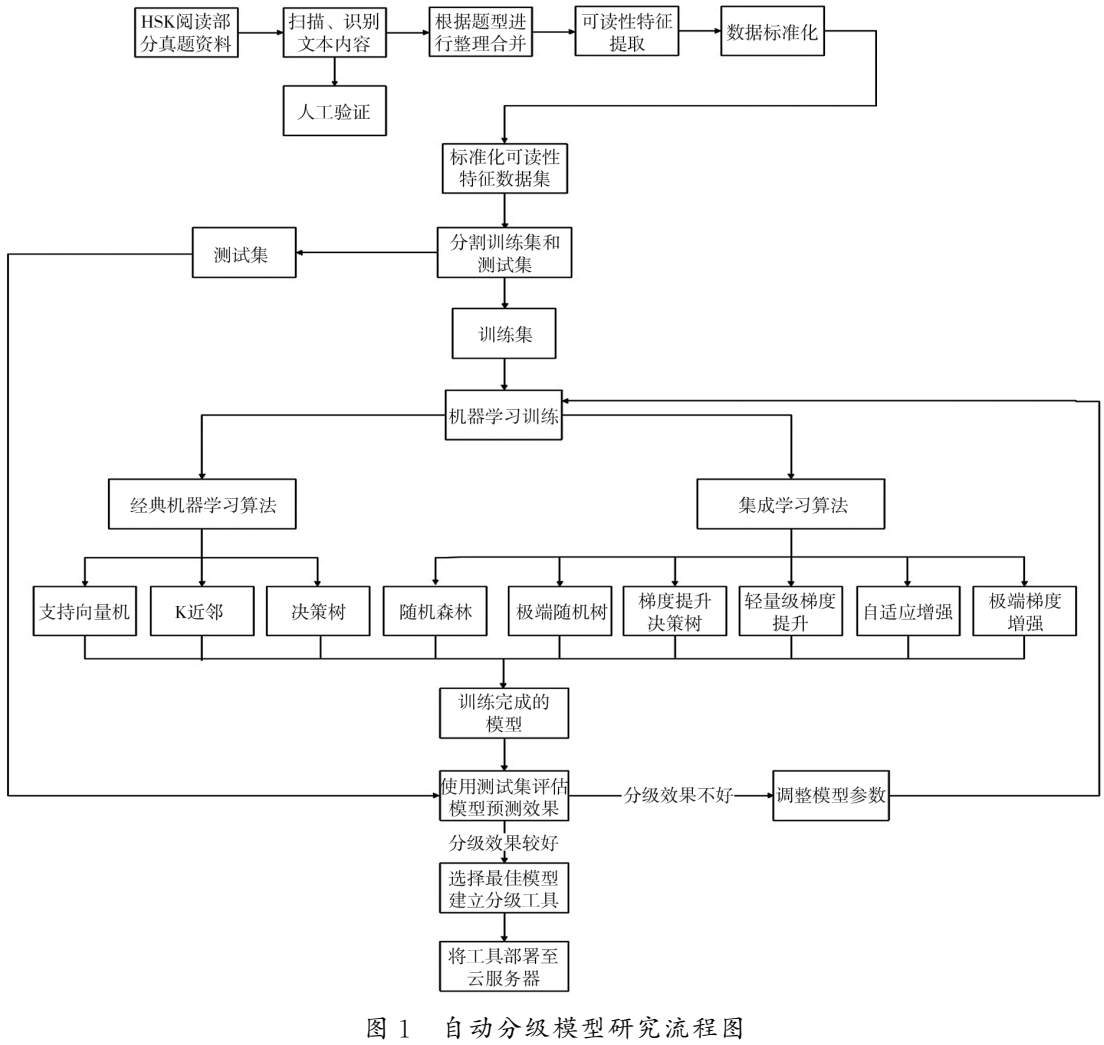

图 1 自动分级模型研究流程图

Fig.1 Flow chart of automatic grading model

1.2.4　文本可读性特征

目前，关于汉语可读性已形成较为成熟的特征集合。本研究从提升模型准确率和兼顾运算性能的角度出发，主要从以下 4 个方面进行特征选择：1）根据以往研究内容和 HSK 阅读文本特点，选取汉字、词汇、句法和篇章 4 个维度共 59 个特征[15]；2）根据《HSK 考试大纲》（以下简称《大纲》）词汇表，制定 1—6 级词汇比例共 6 个特征；3）结合《国际中文教育中文水平等级标准》（GF 0025—2021）（以下简称《标准》），制定 1—9 级汉字和词汇比例等 19 个特征；4）加入 BCC 构建的汉语词频表，该词频表来自 BCC 语料库的报刊、博客、微博和文学频道，共 1 818 656 词。选取特征共计 85 个，如表 2 所示。

表 2　纳入的文本可读性特征

Tab.2　Text readability features included

特征维度	特征名称	特征数量	
		分类	合计
汉字	平均笔画数	1	14
	笔画数标准差	1	
	汉字多样性	1	
	汉字常用度（基于《现代汉语词典》词频表）	1	
	《标准》各级汉字比例（1—9 级）	7	
	《标准》各级手写汉字比例（初、中、高级）	3	
词汇	加权词汇长度	1	32
	词汇类符/形符比	1	
	词汇多样性（UTTR，RTTR）[16]	2	
	形符字词比	1	
	词性复杂度（名词、动词、形容词、副词、数词、量词、助词、介词、实词、虚词、实虚词比）	11	
	《大纲》各级词汇比例（1—6 级）	6	
	《标准》各级词汇比例（1—9 级，初、中、高级）	9	
	词汇常用度（根据 BCC 词频表）	1	
句法	句子长度字级别平均值、最大值	2	20
	句子长度词级别平均值、最大值	2	
	短语类型复杂度（主谓、动宾、动补、双宾、介宾、并列）	6	
	句法结构复杂度（把、被、连、不、没、无、莫、非）字句	8	
	句法语义复杂度语义角色平均值、最大值	2	
篇章	单文本总字数、总词数、总句数	3	19
	单文本指代数（总数、指示代词、疑问代词、第一人称、第二人称、第三人称）	6	
	单文本衔接数（但是、因为、所以、虽然、如果、并且、而且、然后、同时、首先）	10	

为提取和计算上述文本可读性特征，本研究采用 Python 软件下的 Jieba 分词工具进行词语切分、词性标注和词频统计，使用 HanLP 自然语言处理工具进行命名实体识别和句法分析，编写字、词、句、篇 4 个代码模块。同时，结合人工校对方式，构建 HSK 真题阅读文本可读性特征数据集。

2　研究结果

2.1　文本可读性特征数据集

按照前文所述方法收集资料，形成 HSK 真题阅读文本可读性特征数据集，见表 3。所有数据经 K-S 正态分布检验，将符合正态分布的数据采用均数±标准差表示，不符合的数据采用中位数和四分位间距表示。由于篇幅限制，表 3 中仅列出部分特征。

表3 HSK真题阅读文本可读性特征数据集(部分)

Tab.3 HSK reading text readability characteristics data set (part)

序号	特征名称	HSK等级					
		1级	2级	3级	4级	5级	6级
1	平均笔画数	4.66±0.34	5.35±0.32	5.63 (5.27,5.86)	6.02 (5.68,6.23)	6.15 (5.89,6.39)	6.36 (6.18,6.56)
2	汉字多样性	0.68 (0.66,0.72)	0.65 (0.61,0.71)	0.59 (0.55,0.64)	0.57 (0.54,0.62)	0.45±0.05	0.46 (0.41,0.50)
3	汉字常用度	0.29±0.07	0.33±0.07	0.35±0.06	0.40±0.07	0.41±0.05	0.40±0.05
4	加权词汇长度	1.35±0.05	1.39±0.06	1.43±0.07	1.52(1.47,1.56)	1.56±0.07	1.62±0.06
5	词汇类符形符比	0.67 (0.64,0.72)	0.69±0.06	0.64 (0.59,0.68)	0.63 (0.59,0.68)	0.55±0.06	0.58 (0.53,0.61)
6	词汇多样性(UTTR)	17.45 (15.80,20.37)	22.47 (20.58,24.97)	23.04 (19.80,26.09)	25.57 (22.79,28.36)	23.60 (21.31,26.81)	26.55±4.26
7	形符字词比	1.35±0.05	1.39±0.06	1.43±0.07	1.52(1.47,1.56)	1.56±0.07	1.62±0.06
8	词性复杂度_名词	0.12±0.04	0.11±0.03	0.12±0.03	0.13±0.03	0.16±0.04	0.19±0.04
9	词性复杂度_动词	0.14±0.04	0.17±0.03	0.16±0.04	0.18±0.03	0.18±0.03	0.17±0.03
10	《大纲》1级词汇比例	0.14±0.04	0.09±0.03	0.07±0.03	0.07 (0.05,0.08)	0.06±0.02	0.06 (0.04,0.07)
11	《大纲》4级词汇比例	0.07±0.03	0.07±0.03	0.07±0.03	0.19±0.04	0.11±0.03	0.10 (0.09,0.12)
12	《标准》1级词汇比例	0.54±0.07	0.52±0.05	0.44±0.05	0.36±0.05	0.30 (0.27,0.34)	0.25±0.04
13	《标准》4级词汇比例	0 (0,0)	0 (0,0.01)	0.02 (0.01,0.02)	0.04 (0.03,0.06)	0.04 (0.03,0.05)	0.05 (0.04,0.06)
14	《标准》初级(1—3级)词汇比例	0.58±0.06	0.61±0.05	0.61 (0.57,0.65)	0.59 (0.55,0.63)	0.51±0.06	0.44±0.06
15	《标准》高级(7—9级)词汇比例	0 (0,0)	0 (0,0)	0 (0,0)	0.01 (0,0.01)	0.04 (0.03,0.05)	0.05 (0.04,0.07)
16	《标准》1级汉字比例	0.35±0.06	0.38±0.05	0.32±0.05	0.28±0.04	0.26±0.04	0.22±0.03
17	《标准》4级汉字比例	0 (0,0)	0 (0,0)	0 (0,0)	0 (0.00,0.01)	0.01 (0.00,0.01)	0.01 (0.00,0.01)
18	词汇常用度(BCC)	0.026±0.010	0.029±0.009	0.030±0.008	0.036±0.009	0.038±0.008	0.041 (0.034,0.046)
19	句子长度字级别平均值	13.18 (12.00,15.67)	18.14 (15.25,23.00)	23.46 (19.13,29.75)	33.75 (27.79,163.20)	35.08 (30.81,40.24)	37.31 (32.44,42.77)
20	句子长度词级别平均值	9.29 (8.40,11.00)	12.43 (10.75,16.50)	16.00 (12.60,19.61)	21.82 (17.95,96.75)	22.00 (19.38,25.56)	22.74±5.11
21	单文本总字数	63 (55,75)	118 (82,131)	163 (143,233)	228 (200,301)	451 (393,507)	506 (454,643)
22	单文本总词数	52 (43,58)	93 (62,101)	124 (108,160)	158 (140,198)	293 (255,326)	327 (285,402)
23	单文本总句数	5 (4,5)	5 (4,7)	8 (6,11)	8 (6,12)	13 (10,15)	15 (12,19)
24	单文本总指代数	8 (6,9)	11 (8,14)	14 (9,17)	13 (9,16)	18 (12,24)	14 (9,20)

2.2 监督学习模型分级效果

各监督学习模型和Logistic回归的分级效果见表4。由表4可以看出,分级效果最好的是XGBoost模型,准确率为0.913,AUC为0.994,其他3项指标也均位列第一。其余监督学习模型分级准确率均在0.758以上,AUC均在0.917以上。而Logistic回归模型分级性能较监督学习模型有一定的差距,准确率为0.598,AUC为0.857,其他3项指标也均排在末位。

XGBoost模型的混淆矩阵和ROC曲线见图2,图中数值均为5次建模结果的中值。

表 4　各监督学习模型和 Logistic 回归模型的分级效果

Tab. 4　Grading effect of each supervised learning model and Logistic regression

排名	模型	准确率	精确率	召回率	F1-Score	AUC
1	XGBoost	0.913	0.914	0.913	0.913	0.994
2	LGBM	0.911	0.911	0.911	0.910	0.992
3	GBDT	0.902	0.903	0.902	0.902	0.990
4	AdaBoost	0.901	0.904	0.901	0.900	0.991
5	ETC	0.895	0.898	0.895	0.894	0.989
6	RF	0.894	0.896	0.894	0.894	0.987
7	DT	0.812	0.814	0.812	0.811	0.917
8	SVM	0.782	0.786	0.782	0.781	0.939
9	KNN	0.758	0.763	0.758	0.758	0.925
10	Logistic	0.598	0.600	0.598	0.587	0.857

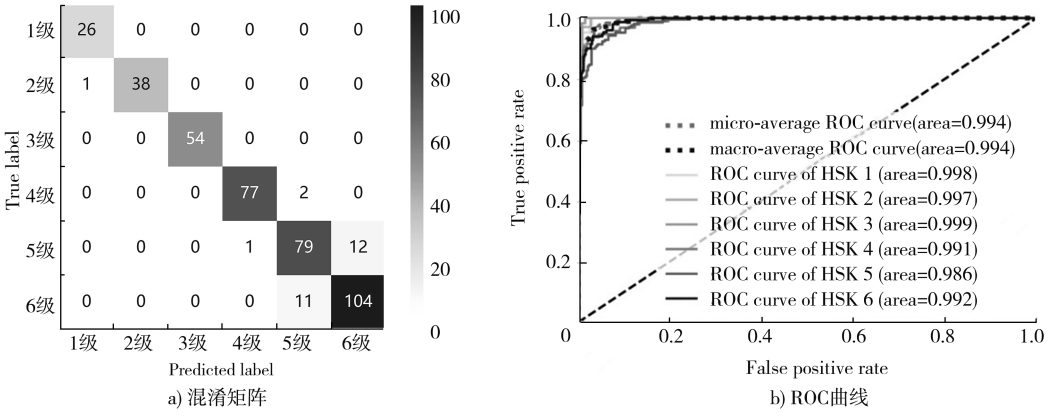

图 2　XGBoost 模型的混淆矩阵和 ROC 曲线

Fig. 2　Confusion matrix and ROC curve for the XGBoost model

2.3　特征重要性

为进一步辨别各特征对分级结果的影响程度，进行特征重要性分析。由于 XGBoost 模型分级效果最好，且具有特征分析功能，故使用该模型进一步计算特征权重数值。图 3 列出了权重排名前 20 的特征。结果表明，与分级最相关的是《大纲》和《标准》部分级别的词汇比例，其中《大纲》6 级词汇比例的重要性明显高于其他特征。

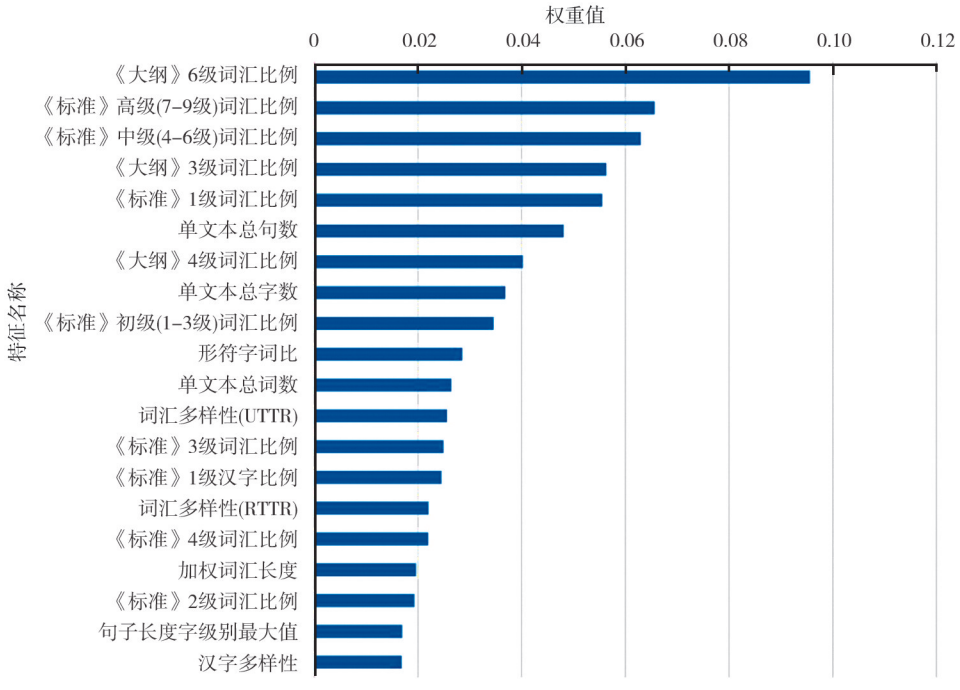

图 3　特征权重排名(前 20 位)

Fig. 3　Feature weight ranking (top 20)

只观察特征权重排名尚无法确定将多少个特征纳入模型可以达到最好的效果，故使用RFE计算最佳特征个数。常用的RFE基础算法包括回归以及SVM，DT和RF等。由于RF在本次实验中得分相对较高，故将其作为基础算法，使用5倍交叉验证计算，得出最佳特征数为21个，如图4所示。

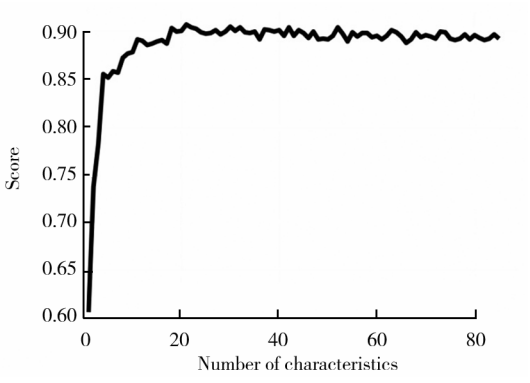

图4 RFE特征数量选择曲线

Fig. 4 RFE feature number selection curve

2.4 模型优化

根据特征重要性和最优特征个数的计算结果，将权重排名前21位的特征纳入各监督学习算法，建模结果显示仍以XGBoost算法的分级效果最好，准确率和AUC分别达到0.919和0.995。其余模型的分级效果也有不同程度的提高。

从特征权重排名可以看出，3项与文本长度相关的指标（单文本总字数、总词数和总句数）与分级结果相关性较高，这与HSK各级别题目的文字量相一致。考虑到用户自选的文本在字词方面的难度不一定与文本长度成正比，为避免文本过长或过短对分级结果的影响，本研究结果呈现为包括和去除文本长度特征2种情况。

在可读性特征集中去除上述3项文本长度特征后，再次使用XGBoost算法和RFE进行建模、特征权重排序和最优特征个数计算。由于去除的特征权重较高，因而XGBoost模型分级效果有所下降，准确率为0.903，AUC为0.990。剩余各特征之间的相对排名较前无明显变化，最佳特征数为25个。对纳入排名前25位的特征再次使用XGBoost算法进行建模，模型分级准确率为0.908，AUC为0.992，分级效果如表5所示。

表5 纳入不同特征后XGBoost模型的分级效果

Tab. 5 Grading effect of XGBoost with different features included

编号	特征选择	特征数量	准确率	精确率	召回率	F1-Score	AUC
1	全部特征	85	0.913	0.914	0.913	0.913	0.994
2	全部特征→特征筛选	21	0.919	0.921	0.919	0.919	0.995
3	全部特征→去除文本长度特征	82	0.903	0.905	0.903	0.903	0.990
4	全部特征→去除文本长度特征→特征筛选	25	0.908	0.909	0.908	0.908	0.992

2.5 在线工具

分别使用表5中的第2和第4项模型建立文本自动分级工具，网址为http://www.hskclassify.online，可通过Web浏览器访问。

3 分析与讨论

3.1 监督学习在HSK阅读文本分级中的应用

本研究中，特征筛选后的XGBoost模型分级准确率达到了0.913，较已有研究[2-3]有了明显提升，其余监督学习模型的准确率也均在0.758以上。与之相比，Logistic回归模型在相同数据条件下的准确率仅为0.598。Logistic回归是一种线性分类器，主要处理二分类问题，并且要求数据必须线性可分，不能有效处理多分类问题或者非线性数据。当特征空间很大时，Logistic回归的性能也会受到明显影响[17]。相比之下，监督学习包含多种类型的算法和技术，具有优秀的计算效能和良好的鲁棒性，可以提升文本分类的准确度和灵活性[18]，能够处理规模较大的数据和任务，如多分类问题、回归问题和聚类问题等[19]。

在监督学习中，Boosting算法是一个比较新的分支，其核心思想是通过迭代方式，不断调整数据的权重分布，使得前一个弱分类器分错的样本在后续模型中得到更多的关注，从而使整体模型更好地对这些困难样本进行分类[20]。在本研究建立的9种监督学习模型中，4种Boosting算法（XGBoost，LGBM，GBDT和AdaBoost）均达到了较好的分级效果，准确率均在0.901以上；2种Bagging算法（ETC，RF）性能稍弱，准确率分别为0.895和0.894；而3种经典模型（DT，SVM和KNN）准确率分别为0.812，0.782和0.758，与上述模型相比有一定的差距。整体来看，Boosting算法在HSK阅读文本分析方面具有优势。这表明在遇到汉语文本可读性数据分析问题时，应当纳入监督学习特别是Boosting算法，并与其他算法进行对比，择优选用，以达到更好的分析效果，使研究结果更具指导性和针对性。

采用准确率、AUC 等多项指标评价各模型的分级效果，可以全面了解模型性能。准确率是文本可读性研究中使用较多的一个指标，指的是被正确分类的样本数与总样本数的比值。如果一个分类模型的准确率高，说明该模型能够很好地将不同类别的样本区分开。但在样本不均衡的情况下，准确率可能无法准确反映模型性能[21]。AUC 是一个在监督学习领域更加常用的评估指标，衡量模型在所有可能的分类阈值下的表现，可以反映模型对多类别的整体排序能力。AUC 同时考量对正例和负例的区分，在样本不均衡的情况下，依然能够合理评估模型性能[22]。本研究中，HSK 不同级别的文本字数和所生成的文本条数差距较大：1 级为 6 451 字，102 条文本；6 级为 191 669 字，352 条文本。因此，除了准确率等指标，本研究还采用 AUC 作为评估标准。在 9 种监督学习算法中，XGBoost 算法的 AUC 值最高，为 0.994，表明该算法具有较高的分级性能和实用价值。

3.2　HSK 阅读文本分级模型的启示

语言的本质是词汇和语法的组合，通过考查词汇的掌握情况，可以更准确地评估语言水平和实际应用能力。本研究主要采用《大纲》1—6 级和《标准》1—9 级词汇表。从特征筛选结果来看，《大纲》和《标准》不同级别的词汇比例在前 10 项中占据了 7 项，在去除 3 个文本长度特征后更是占据了 9 项，且分级准确率较高。这表明不同难度等级的词汇是影响 HSK 阅读文本分级的最主要因素。因此，在针对 HSK 的研究中，应当特别重视对词汇的学习和使用。

《标准》的制定与《大纲》关系密切。在词汇量方面，《大纲》词汇总量为 5 000 个，《标准》以《大纲》为基础进行了扩充和更新，词汇总量为 11 092 个。研究显示，《大纲》中的 4 392 个词汇被收入《标准》中。本研究特征筛选结果表明，《大纲》和《标准》的各级词汇比例对于 HSK 分级的影响程度基本等同，这与两者共有词汇较多的现象相符。为了判断本研究成果对 HSK 变化的适应能力，在数据集中去除了《大纲》1—6 级词汇比例这 6 个特征，再次进行模型训练和特征筛选。结果显示，分级准确率仍可达 0.881，《标准》各级词汇比例在特征权重排名前 10 项中占据了 6 项。

文本长度也是影响 HSK 分级的重要因素。在 HSK 1—6 级中，每份阅读部分的平均字数为 190～7 114，HSK 6 级的阅读大题单篇文字量可达 1 000 字。这提示在遴选 HSK 阅读材料时，应注意文本长度与难度之间的相关性，适当增加单篇千字以上的长文本阅读训练。

本研究建立了在线分级工具，可对自选文本进行相应的 HSK 分级，对于介于两级之间的文本，可显示属于每一级的概率数值，便于用户综合判断。在具体操作中，根据 HSK 各级阅读真题的字数情况，建议输入文本的字数在 50～1 000 之间。对于过长或过短的文本，应以去除文本长度特征后的分级结果为主。

4　结　语

1）基于监督学习的 HSK 阅读材料自动分级模型研究结果显示，各级词汇比例是影响文本分级的主要因素。经参数优化和特征筛选，XGBoost 算法在各监督学习模型中的分级效果最好，准确率为 0.919。在 HSK 阅读文本分析方面，监督学习较 Logistic 回归有明显优势，其中又以 Boosting 表现最佳。

2）本研究建立的分级模型和在线工具能够以较高的准确率对自选文本进行 HSK 等级分类，帮助用户有针对性地遴选，提高学习效率。

后续研究将根据 HSK 阅读文本在词汇和其他可读性特征上的变化，调整模型参数和特征权重，及时更新研究成果。同时，紧跟监督学习领域的发展，对所建立的模型和在线工具不断进行优化，添加更多功能，达到更好的使用效果。

参考文献/References：

[1] PENG Yue，YAN Wei，CHENG Liying. HSK：A multi-level，multi-purpose proficiency test[J]. Language Testing，2021，38(2)：326-337.

[2] 江新，宋冰冰，姜悦，等. 汉语水平考试(HSK)阅读测试文本的可读性分析[J]. 中国考试，2020(12)：30-37.
JIANG Xin，SONG Bingbing，JIANG Yue，et al. A study on the readability of reading test texts in Chinese proficiency test(HSK)[J]. China Examinations，2020(12)：30-37.

[3] 杜月明，王亚敏，王蕾. 汉语水平考试(HSK)阅读文本可读性自动评估研究[J]. 语言文字应用，2022(3)：73-86.
DU Yueming，WANG Yamin，WANG Lei. A study on the automatic text readability assessment of reading texts in Hanyu Shuiping Kaoshi(HSK)[J]. Applied Linguistics，2022(3)：73-86.

[4] 张庆翔,张莹. 国际中文教育的文本可读性研究回顾[J]. 现代语文,2022(10):89-95.
ZHANG Qingxiang,ZHANG Ying. A review of research on the readability of international Chinese language education texts[J]. Modern Chinese,2022(10):89-95.

[5] 夏菁,孙未未. 多向度计量语体特征下的对外汉语教材可读性自动评估研究[J]. 华中学术,2020,13(2):181-193.

[6] 孙未未,夏菁,曾致中. 基于回归模型的对外汉语阅读材料的可读性自动评估研究[J]. 中国教育信息化,2018(15):67-74.

[7] 杨文娣,曾致中. 基于随机森林算法的对外汉语文本可读性评估[J]. 中国教育信息化,2019(14):89-96.

[8] 许琦,姚锦江. 基于特征提取和机器学习的数据可视化模型构建研究[J]. 自动化与仪器仪表,2023(12):38-41.
XU Qi,YAO Jinjiang. Research on data and information visualization model construction based on feature extraction and machine learning [J]. Automation & Instrumentation,2023(12):38-41.

[9] 盛雪晨. 基于分布式机器学习的文本分类模型研究[D]. 南京:南京邮电大学,2023.
SHENG Xuechen. Text Classification Model Basedon Distributed Machine Learning[D]. Nanjing:Nanjing University of Posts and Telecommunications,2023.

[10] 李艳,朱倩倩,董秀萍. 基于机器学习模型的客服短文本分类技术研究[J]. 现代计算机,2023,29(15):64-68.
LI Yan,ZHU Qianqian,DONG Xiuping. Research on short text classification technology of customer service based on machine learning model[J]. Modern Computer,2023,29(15):64-68.

[11] 刘滨,詹世源,刘宇,等. 基于密度 Canopy 的评论文本主题识别方法[J]. 河北科技大学学报,2023,44(5):493-501.
LIU Bin,ZHAN Shiyuan,LIU Yu,et al. Topic recognition method of comment text based on density Canopy[J]. Journal of Hebei University of Science and Technology,2023,44(5):493-501.

[12] 刘滨. 分布式数据挖掘综述[J]. 河北科技大学学报,2014,35(1):79-90.
LIU Bin. Survey on distributed data mining[J]. Journal of Hebei University of Science and Technology,2014,35(1):79-90.

[13] 于卫红. 多类别文本分类方法比较研究[J]. 计算机技术与发展,2022,32(1):54-60.
YU Weihong. Study on comparison of multi-class text classification methods[J]. Computer Technology and Development,2022,32(1):54-60.

[14] GONZÁLEZS S, GARCÍA S,SER J D,et al. A practical tutorial on bagging & boosting based ensembles for machine learning:Algorithms,software tools,performance study,practical perspectives & opportunities[J]. Information Fusion,2020(64):205-237.

[15] 吴思远,蔡建永,于东,等. 文本可读性的自动分析研究综述[J]. 中文信息学报,2018,32(12):1-10.
WU Siyuan,CAI Jianyong,YU Dong,et al. A survey on the automatic text readability measures[J]. Journal of Chinese Information Processing,2018,32(12):1-10.

[16] DU Yueming. The relationship of lexical richness to the quality of CSL writings[C]//Lecture Notes in Computer Science. Cham:Springer,2023:116-131.

[17] WESTREICH D, LESSLER J, FUNK M J. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression[J]. Journal of Clinical Epidemiology,2010,63(8):826-833.

[18] 杨晓哲,王晴晴,蒋佳龙. 基于人工智能的课堂师生对话分析:IRE 的自动分类与分水平构建[J]. 电化教育研究,2023,44(10):79-86.
YANG Xiaozhe,WANG Qingqing,JIANG Jialong. Analysis of classroom teacher-student dialogue based on artificial intelligence:Automatic classification and sub-level construction of IRE[J]. E-education Research,2023,44(10):79-86.

[19] FEURER M,KLEIN A,EGGENSPERGER K,et al. Efficient and robust automated machine learning[J]. Advances in Neural Information Processing Systems,2016,28:2944-2952.

[20] MAYR A,BINDER H,GEFELLER O,et al. The evolution of boosting algorithms[J]. Methods of Information in Medicine,2014,53(6):419-427.

[21] ABDELRAHMAN S M A,ABRAHAM A. A review of class imbalance problem[J]. Journal of Network and Innovative Computing,2013,1:332-340.

[22] LINGC X,HUANG J,ZHANG H. AUC:A better measure than accuracy in comparing learning algorithms[C]//Advances in Artificial Intelligence. Berlin:Springer,2003:329-341.