本文引用: 魏方志, 潘承丹, 宋逸天, 庄燕苹, 张 绚, 曾旻昱, 贾晓康, 宫爱民. 基于XGBoost算法的系统性红斑狼疮中医证型判别模型研究[J]. 湖南中医药大学学报, 2024, 44(12): 2286-2293.

基于XGBoost算法的系统性红斑狼疮中医证型判别模型研究

魏方志[1,2],潘承丹[1],宋逸天[1],庄燕苹[1],张 绚[1],曾旻昱[1],贾晓康[1],宫爱民[1]*

1.海南医科大学(海南医学科学院)中医学院,海南 海口 571199;2.博鳌一龄生命养护中心,海南 琼海 571400

〔摘要〕**目的** 通过XGBoost算法构建系统性红斑狼疮(systemic lupus erythematosus, SLE)中医证型判别模型,探索XGBoost模型用于证型分类的可行性。**方法** 通过问卷调查法,收集符合标准的病例,建立SLE数据集。通过XGBoost算法构建SLE中医证型判别模型,采用随机森林(random forest, RF)算法作为对照,比较两种算法的准确性。**结果** 本研究共纳入400例SLE患者,其中男性33例,女性367例。SLE患者排名前3的中医证型为:脾肾阳虚证、阴虚内热证和风湿热痹证,XGBoost算法模型分类指标和性能曲线评分总体优于RF算法。**结论** XGBoost算法用于证候建模准确度较高,可用于证候研究中的分类研究。

〔关键词〕系统性红斑狼疮;XGBoost算法;随机森林算法;中医证候

〔中图分类号〕R259 〔文献标志码〕A 〔文章编号〕doi:10.3969/j.issn.1674-070X.2024.12.021

Chinese medicine pattern differentiation model for systemic lupus erythematosus based on XGBoost algorithm

WEI Fangzhi[1,2], *PAN Chengdan*[1], *SONG Yitian*[1], *ZHUANG Yanping*[1], *ZHANG Xuan*[1], *ZENG Minyu*[1], *JIA Xiaokang*[1], *GONG Aimin*[1]*

1. School of Chinese Medicine, Hainan Medical University (Hainan Academy of Medical Sciences), Haikou, Hainan 571199, China; 2. Bo'ao Yiling Life Care Center, Qionghai, Hainan 571400, *China*

〔**Abstract**〕**Objective** To construct a Chinese medicine (CM) pattern differentiation model for systemic lupus erythematosus (SLE) using the XGBoost algorithm and explore the feasibility of applying the XGBoost model for CM pattern classification. **Methods** Eligible cases were collected through a questionnaire survey to establish a SLE dataset. An XGBoost-based SLE CM pattern differentiation model was developed, and the random forest (RF) algorithm was used as a control for accuracy comparison. **Results** A total of 400 SLE patients were included in this study, including 33 males and 367 females. The top three CM patterns for SLE patients were yang deficiency of the spleen and kidney pattern, yin deficiency-induced internal heat pattern, and wind dampness and heat impediment pattern. The classification indicators and performance curve scores of the XGBoost algorithm model were overall superior to those of the RF algorithm. **Conclusion** XGBoost algorithm demonstrates high accuracy in CM pattern modeling and can be used for classification research in CM pattern studies.

〔**Keywords**〕systemic lupus erythematosus; XGBoost algorithm; random forest algorithm; Chinese medicine pattern

〔收稿日期〕2024-05-22

〔基金项目〕国家自然科学基金项目(30109065)。

〔通信作者〕* 宫爱民,男,博士,教授,博士研究生导师,E-mail:422789075@qq.com。

系统性红斑狼疮(systemic lupus erythematosus, SLE)是一种涉及多器官、多组织的自身免疫疾病。我国患病率约为 1/10 000,是西方国家的 2 倍[1]。SLE 与中医学“红蝴蝶疮”“红斑痹”“阴阳毒”等类似。SLE 初期多热证,后期多阴虚证或阳虚证,而瘀血始终贯穿其中。中医药能从整体角度调理 SLE 患者,不仅能改善 SLE 患者的症状,还能减少西药的毒副作用[2-4]。辨证论治是中医处方治疗的核心,正确的辨证对 SLE 治疗十分关键。但中医辨证缺乏特异性指标,证型判读缺乏客观性、重复性和系统性。

近年来,人工智能在辅助中医诊断和治疗等方面发挥巨大潜力价值,已成为中医证候学客观化研究的重要方法之一。目前,中医药领域常用的机器学习算法有聚类分析、贝叶斯网络、支持向量机、决策树和随机森林(random forest, RF)等[5],这些算法都存在一些弊端,如贝叶斯网络需要得到先验概率,决策树较容易过拟合且难以寻找到最佳的树,支持向量机选择适当的核函数比较困难,RF 处理小样本效果欠佳。

XGBoost 算法可以对小样本、半定量的中医数据进行高效的处理,与中医辨证思维有一定契合度,在证候分类中具有潜在价值。本研究拟以 SLE 中医证候判别模型为切入点,运用 XGBoost 算法构建 SLE 证型判别模型,同时引入 RF 算法作为对照,对比两种模型的性能。进一步通过参数调优,根据优化模型的分类指标和性能曲线评分,筛选出更适合本数据集的算法模型。

1 材料与方法

1.1 研究对象

本研究病例主要来源于 2020 年 7 月至 2022 年 1 月在海南医学院第一附属医院和海南医学院附属海南医院门诊和住院部就诊的SLE 患者。本研究经海南医学院伦理委员会批准,批准号为 HYLL-2022-239。

1.2 诊断标准

1.2.1 SLE 西医诊断标准　参照 2019 年欧洲抗风湿病联盟和美国风湿病协会修订的 SLE 诊断标准[6]。(1)抗核抗体阳性:将抗核抗体至少一次阳性列为强制性准入标准。(2)临床指标和免疫学指标:7 项临床指标(全身状态、血液学、神经心理学、皮肤黏膜、浆膜、肌肉骨骼、肾脏改变)和 3 个免疫学指标(抗磷脂抗体、补体、SLE 特异性抗体)。满足至少一项临床指标,免疫学指标积分≥10 分。

1.2.2 SLE 中医证型诊断标准　参照 2002 年《中药新药临床研究指导原则》[7]中的标准将 SLE 分为热毒炽盛证、阴虚内热证、脾肾阳虚证、肝肾阴虚证、瘀热痹阻证、风湿热痹证和气血两虚证 7 种证型。

1.3 纳入、排除及剔除标准

1.3.1 纳入标准　(1)符合 SLE 西医诊断标准;(2)自愿参与本研究;(3)年龄为 18~80 岁。

1.3.2 排除标准　(1)合并脑、心、肝、肾和造血系统等其他严重疾病者;(2)合并精神、神经疾病者;(3)妊娠或哺乳期妇女;(4)不愿配合本试验者。

1.4 样本量的估算

多因素研究主要是根据研究因素来决定样本量大小,在专家咨询、文献研究和前期小样本验证下,设置 80 个研究因素。多因素研究的样本量通常使用简单估算法,即样本量至少为研究因素的 5~10 倍[8],因此,本研究共纳入 400 例样本。

1.5 SLE 中医调查表的制定

本研究通过 2004 年《中医症状鉴别学》[9]、2005 年《中医临床常见症状术语规范》[10]和“十三五”规划教材《中医诊断学》[11]进行中医四诊术语规范化描述。在专家的指导下制定出《SLE 中医调查表》,其中症状、体征按程度分为无、轻、中、重,分别记为 0、2、4、6 分,舌脉按有、无分别记为 2、0 分。

1.6 研究对象的筛选方法

对于符合标准的 SLE 受试者,由 3 位中医医师进行基本信息、四诊信息的收集和中医证型的判读,需同时满足 2 位及以上医师判读结果一致方可纳入研究。

1.7 数据预处理及数据集的建立

将 SLE 四诊信息条目按照 f1~f80 依次编码,热毒炽盛证编码为 0,阴虚内热证编码为 1,脾肾阳虚证编码为 2,肝肾阴虚证编码为 3,风湿热痹证编码为 4,瘀热痹阻证编码为 5,气血两虚证编码为 6。将采集的 SLE 患者信息双人背靠背录入,检查数据的完整性和一致性,建立适用于本研究的 SLE 数据集。详见表 1。

表 1 SLE 数据集

Table 1 SLE dataset

f1	f2	f3	f4	f5	f6	……	f79	f80	证型
0	0	0	2	0	0	……	0	2	2
0	0	2	0	2	0	……	0	0	1
0	0	0	0	0	0	……	0	0	4
0	0	0	4	0	0	……	0	0	4
0	0	2	0	2	0	……	0	0	1
0	0	0	0	0	0	……	0	0	2
0	0	0	0	0	0	……	0	0	1
0	0	2	0	0	0	……	0	0	3
0	0	0	0	0	0	……	0	0	2
0	0	0	0	0	0	……	0	0	1
0	0	0	0	0	0	……	0	0	3
0	0	2	0	0	0	……	2	0	1
0	0	0	0	2	0	……	0	0	1
0	0	0	0	0	0	……	0	0	2
0	0	0	0	0	0	……	0	2	2

1.8 算法初步模型的建立及优化

1.8.1 RF 算法原理 RF 是一种包含多棵决策树的集成学习算法，输出结果由输出类别的平均数或众数而定。其算法原理是基于集成学习的装袋法，装袋法是通过构建多个相互独立的弱分类器，根据其预测结果来评估弱分类器的效果。在分类问题中，其预测步骤如下：首先使用随机构建的分类器测试数据结果，然后计算每种预测分类结果的票数，最后将获得票数最高的分类结果视为最终预测结果。

1.8.2 XGBoost 算法原理 XGBoost 是一种基于强分类器的增强集成学习算法，输出结果由强分类器结果而定。该算法原理是基于集成学习的提升法，提升法是将多个弱分类器集合成一个强分类器[12]。其算法步骤为：

(1)初始目标函数：

$$Ob_j = \sum_{i=1}^{n} l\left(y_i, y_i\right) + \sum_{k=1}^{K} \Omega(f_k) \quad ①$$

初始目标函数包含两个部分：第一部分是模型的训练误差，第二部分是正则化项，正则化项是由 K 棵树的正则化项相加而来。

(2)改写目标函数：

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad ②$$

(3)目标函数泰勒二阶展开：

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, y_i^{(t-1)} + g_i f_t(x_i)\right) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad ③$$

其中，g_i 为一阶导数，h_i 为二阶导数。

(4)最优目标函数：

$$Ob_j = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad ④$$

该过程可总结为：XGBoost 首先根据数据集生成一棵树，得到初始目标函数，并不断对树进行添加，形成新的目标函数，并用新的目标函数结果对上次的预测残差进行拟合。在所有数据集训练结束后，可得到 n 棵树。根据样本特征，找到每棵树的叶子节点分数，将全部叶子节点分数累加即为样本的预测结果。

1.8.3 计算机配置信息 CPU：Intel core i7-10700K CPU @ 3.8GHz 处理器；显卡：NVIDIA GeForce RTX 2070 SUPER(8G/微星)；内存：金士顿 DDR4 3200MHz 32G。

1.8.4 主要软件和包 主要软件：Anaconda 3 和 Python 3.9.7。主要包：Numpy、Panda、Matplotlib、scikit-learn 和 XGBoost。

1.8.5 算法建模的流程 (1)导入相应的包；(2)导入 SLE 数据集；(3) 分裂 SLE 数据集的特征和目标值；(4)以 7:3 的比例将 SLE 数据集拆分成训练集和测试集；(5)XGBoost 建模预测；(6)RF 建模预测。

1.8.6 模型的优化 初步建模后的模型通常不是最优模型，需要根据数据特征和任务目标进一步调整相关算法的重要参数，使构建的模型更准确和稳定，以符合临床使用要求。本研究运用网格搜索法和 3 折交叉验证分数筛选XGBoost 和 RF 算法参数的最优值，并结合预测准确率验证最优值的可靠性，完成对算法模型的优化。在 XGBoost 中，n_estimators 代表最大迭代次数，eta 代表学习率，max_depth 代表树的最大深度，min_child_weight 代表最小叶子节点权重和，gamma 代表节点分裂所需的最小损失下降值，subsample 代表随机采样的比例，colsample_bytree 代表随机采样占总样本的比例。在 RF 中，n_estimators 和 max_depth 与 XGBoost 含义相同，max_features 代表最大特征数，min_samples_leaf 代表叶子节点所需的最少样本数，min_samples_split 代表节点划分需要的最小样本数，criterion 代表分裂标准。XGBoost 和 RF 重要参数取值范围见表 2，按表中参数顺序依次筛选参数最佳值。

表 2　参数取值表

Table 2　Parameter value table

分类器	XGBoost	RF
参数取值范围	n_estimators：1~400	n_estimators：1~200
	eta：0.01~0.51	max_depth：1~20
	max_depth：2~20	max_features：5~30
	min_child_weight：1~6	min_samples_leaf：1~10
	gamma：0~0.6	min_samples_split：2~20
	subsample：0.5~1	criterion：[′gini′，′entropy′]
	colsample_bytree：0.5~1	—

注：n_estimators 初次步长为 10，寻找最佳值时步长为 1，eta 步长为 0.01，subsample、gamma、colsample_bytree 步长为 0.1，其他参数步长均为 1。

1.9　模型的评价

为了评价各训练模型的表现，本研究基于分类指标和性能曲线比较不同模型的性能，其中常用的分类指标有：交叉验证分数、准确率、平均准确率、精准率、召回率、F1 值、科恩卡帕分数、宏平均（以下简称“宏”）和微宏平均（以下简称“微”）；性能曲线主要有：ROC 曲线、PR 曲线和学习曲线。交叉验证是将数据集划分为较小子集的用于评估模型性能的方法，本次实验采用 3 折交叉验证平均分数进行模型性能评估。

2　结果

2.1　基本资料

2.1.1　性别　本研究共纳入 400 例 SLE 患者，男性患者 33 例，女性患者 367 例，男性占总人数的8.3%，女性占总人数的 91.7%，男性:女性=1:11.1。

2.1.2　年龄　在 400 例 SLE 患者中，患者年龄 18~79（35.78±13.55）岁。青年（18 岁≤年龄<45 岁）有 290 人，占 72.5%。青年男性 25 人，占总数的 6.25%；青年女性 265 人，占总数的 66.25%。中年（45 岁≤年龄<60 岁）有 84 人，占 21.0%。中年男性 5 人，占总数的 1.25%；中年女性 79 人，占总数的 19.75%。老年（年龄≥60 岁）有 26 人，占 6.5%。老年男性 3人，占总数的 0.75%；老年女性 23 人，占总数的 5.75%。

2.2　中医四诊信息结果

400 例 SLE 患者中医四诊信息排序表见表 3。该表将 80 项中医四诊信息按照频数、频率高低进行降序排列。由表 3 可知，有 19 项中医四诊信息出现频率高于 20%，分别是：舌苔黄、发热、皮肤红斑、关节固定性疼痛、神疲乏力、齿痕舌、舌红、舌淡白、脱发、舌淡红、脉数、舌苔白、舌苔少或无、水肿、脉弱、舌苔腻、纳差、皮疹和脉细。

2.3　中医证型结果

在本次调查的 400 例 SLE 患者中，最常见的中医证型为脾肾阳虚证（110 人，占 27.5%）。其次分别是阴虚内热证（25.3%）、风湿热痹证（15.8%）、热毒炽盛证（11.0%）、瘀热痹阻证（9.8%）、气血两虚证（6.3%）、肝肾阴虚证（4.5%）。详见表 4。

2.4　算法调参结果

XGBoost 重要参数调整最终结果为：n_estimators=40，eta=0.3，subsample=0.5，min_child_weight=1，colsample_bytree=1，objective=multi:softmax，num_class=7，random_state=420。RF 重要参数调整最终结果为：n_estimators=51，max_depth=19，max_features=14，min_samples_leaf=2，min_samples_split=2，criterion=gini，random_state=420，其他未提及参数均为默认值最佳。详见图 1。

2.5　模型评价与验证

2.5.1　两种算法分类指标结果　XGBoost 算法模型整体的准确率、3 折交叉验证分数、平衡准确率、科恩卡帕系数、宏精准率、加权精准率、宏召回率、加权召回率分别为 78.3%、73.9%、63.5%、72.1%、71.0%、79.0%、64.0%、78.0%。RF 算法模型对应指标分数分别为 74.2%、74.3%、57.9%、66.6%、66.0%、74.0%、58.0%、74.0%。XGBoost 算法模型各证型的精准率分别为 89%、86%、64%、0%、89%、71%、100%，RF 算法模型各证型的精准率分别为 90%、86%、60%、0%、80%、67%、75%。详见表 5—6。

2.5.2　两种算法性能曲线结果　XGBoost 和 RF 算法模型的 ROC 曲线见图 2 和图 3。由图可知，XGBoost 模型在微、宏和各证型的 AUC 值分别为 0.93、0.90、0.90、0.97、0.90、0.79、0.88、0.91 和 0.95；RF 对应类型的 AUC 值分别为 0.92、0.88、0.83、0.96、0.92、0.75、0.88、0.89 和 0.90。XGBoost 和 RF 算法模型的 PR 曲线见图 4。由图可知，XGBoost 模型微、宏 AP 值分别为 0.78 和 0.73；RF 模型微、宏 AP 值分别为 0.78 和 0.74。通过上述分类指标和性能曲线对模型进一步评价及验证，可以发现 XGBoost 模型整体表现优于 RF 模型。

表 3 SLE 患者中医四诊信息排序表

Table 3 Ranking table of CM four diagnostic information of SLE patients

排序	中医四诊信息	频数/人	频率	排序	中医四诊信息	频数/人	频率
1	舌苔黄 *	191	47.8%	41	舌质紫暗,或有瘀斑	44	11.0%
2	发热 *	147	36.8%	42	脉滑	44	11.0%
3	皮肤红斑 *	140	35.0%	43	盗汗	44	11.0%
4	关节固定性疼痛 *	138	34.5%	44	便溏	38	9.5%
5	神疲乏力 *	136	34.0%	45	畏光	37	9.3%
6	齿痕舌 *	134	33.5%	46	关节变形	37	9.3%
7	舌红 *	133	33.3%	47	健忘	35	8.8%
8	舌淡白 *	131	32.8%	48	肌肤甲错	33	8.3%
9	脱发 *	121	30.3%	49	脉弦	32	8.0%
10	舌淡红 *	120	30.0%	50	脉濡	32	8.0%
11	脉数 *	120	30.0%	51	胸痛	31	7.8%
12	舌苔白 *	119	29.8%	52	肌肉疼痛	31	7.8%
13	舌苔少或无 *	107	26.8%	53	手足心热	30	7.5%
14	水肿 *	102	25.5%	54	神昏	29	7.3%
15	脉弱 *	94	23.5%	55	脉涩	29	7.3%
16	舌苔腻 *	86	21.5%	56	咽痛	27	6.8%
17	纳差 *	86	21.5%	57	烦躁	27	6.8%
18	皮疹 *	82	20.5%	58	视物模糊	26	6.5%
19	脉细 *	81	20.3%	59	月经不调或闭经	24	6.0%
20	咳嗽	74	18.5%	60	脉洪	24	6.0%
21	舌瘦	73	18.3%	61	口黏腻	24	6.0%
22	舌苔滑	72	18.0%	62	心悸	22	5.5%
23	失眠	67	16.8%	63	尿少	22	5.5%
24	口疮	63	15.8%	64	腰痛	21	5.3%
25	裂纹舌	60	15.0%	65	脉缓	20	5.0%
26	腹胀	60	15.0%	66	脉浮	20	5.0%
27	眩晕	58	14.5%	67	舌苔厚	18	4.5%
28	气短	58	14.5%	68	自汗	17	4.3%
29	关节肿胀	58	14.5%	69	面色无华	17	4.3%
30	潮热	57	14.3%	70	口苦	17	4.3%
31	畏寒肢冷	56	14.0%	71	口渴喜冷饮	17	4.3%
32	舌胖大	56	14.0%	72	恶风	16	4.0%
33	胸闷	55	13.8%	73	便秘	16	4.0%
34	晨僵	55	13.8%	74	脉沉	15	3.8%
35	肢体麻木	52	13.0%	75	尿黄	13	3.3%
36	关节游走性疼痛	50	12.5%	76	口渴喜热饮	10	2.5%
37	腰膝酸软	47	11.8%	77	耳鸣	10	2.5%
38	舌苔薄	46	11.5%	78	足跟痛	9	2.3%
39	咽干	45	11.3%	79	尿频	9	2.3%
40	肢体困重	44	11.0%	80	口渴不欲饮	4	1.0%

注:* 为频率高于 20%的中医四诊信息。

3 讨论

在 400 例 SLE 患者中,男性 33 例,女性 367 例,男:女=1:11.1。男性患病率显著低于女性,与我国 SLE 患者男女比为 1:7~1:13 的报道结果相符[13]。SLE 发病年龄以青年患者居多,占总人数的 92%,其中青年女性占比高达总人数的 66.25%,这与 SLE 以育龄期(20~40 岁)女性多见的报道结果一致[14]。

对中医四诊信息结果分析发现,频率高于 20%的四诊信息与 2002 年《中药新药临床研究指导原

表 4　SLE 患者中医证型分布表

Table 4　Distribution table of CM patterns in SLE patients

排序	中医证型	频数/人	频率
1	脾肾阳虚证	110	27.5%
2	阴虚内热证	101	25.3%
3	风湿热痹证	63	15.8%
4	热毒炽盛证	44	11.0%
5	瘀热痹阻证	39	9.8%
6	气血两虚证	25	6.3%
7	肝肾阴虚证	18	4.5%

则》[7]上的 SLE 常见四诊信息相符，也初步反映了海口地区 SLE 患者常见的症状、体征和舌脉象等中医四诊信息。

在 400 例 SLE 患者中，排行前 3 的证型为脾肾阳虚证、阴虚内热证和风湿热痹证。海口地区 SLE 患者阴虚内热证发病率较高，“瘀热”也较为常见，这与既往研究结果相符[15]。此外，本地区 SLE 患者还有脾肾阳虚证高发、病程多“夹湿”的特点，这可能与 SLE 病机演变和海口地区的地域、气候有关。海口地处热带北缘，气候炎热多湿，居民又嗜好生冷，损伤脾阳，脾虚生湿，因此“夹湿”患者较为多见。研究结果也表明，海南地区患者具有脾虚夹湿的特点[16-18]。因此，研究不同地域相同疾病的中医证候特点，有利于丰富不同地区的中医证候学资料，进一步指导“因地制宜”的治疗方案。

XGBoost 和 RF 算法控制参数较多，初步建立的模型往往需要调参才能得到更准确、更稳定的模型。调参可以控制模型复杂度和泛化误差大小，模型复杂度的高或低会导致模型过拟合或欠拟合。只有方差和偏差最小时，模型才能达到复杂度最佳、泛化误差最小和预测准确率最高。XGBoost 和 RF 模型均是复杂度高的模型，在本数据中两者模型均存在过拟合，因此，调参目标均是降低模型复杂度和方差，防止过拟合。

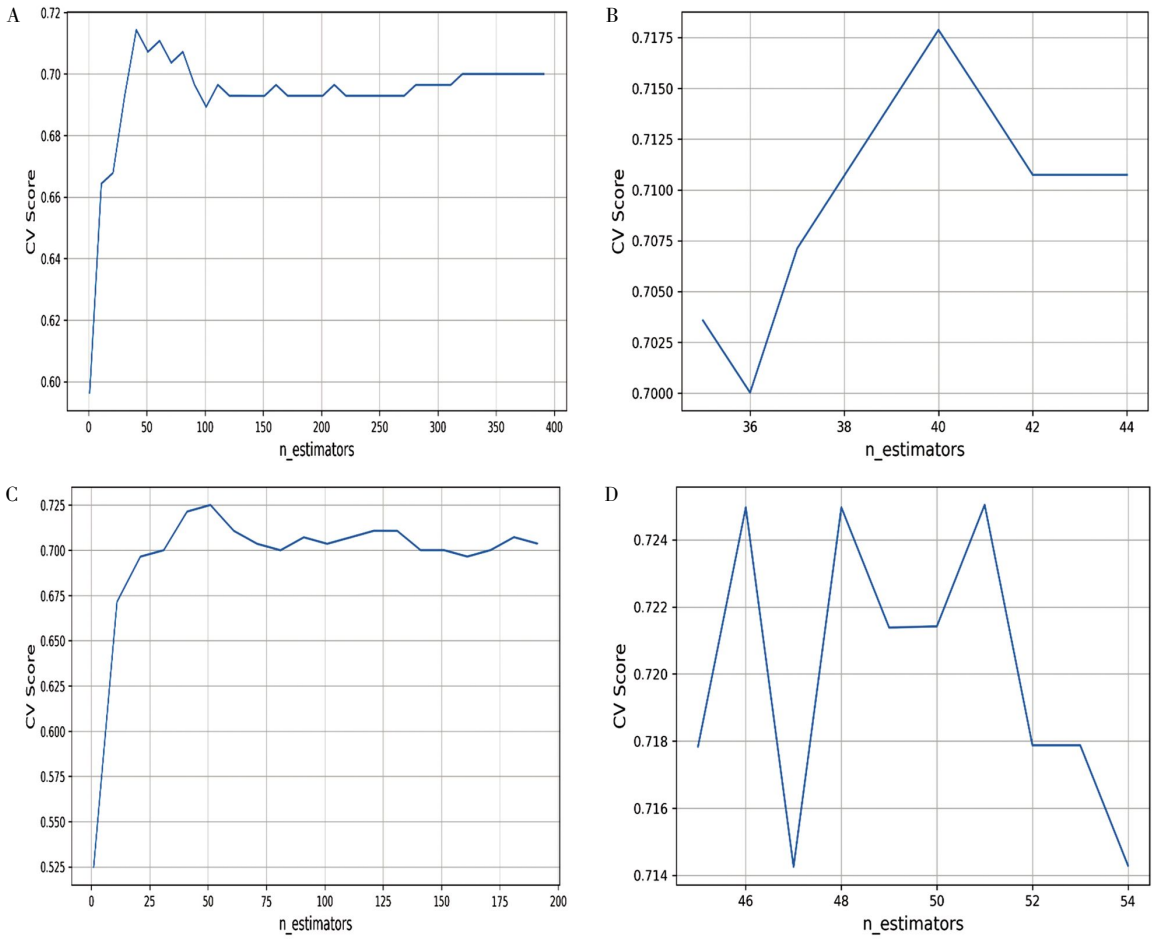

图 1　n_estimators 学习曲线图

Fig.1　Learning curve of n_estimators

注：A. XGBoost 的 n_estimators 学习曲线图；B. XGBoost 最佳 n_estimators 图；C. RF 的 n_estimators 学习曲线图；D. RF 最佳 n_estimators 图。横坐标为最大迭代次数，纵坐标为 3 折交叉验证分数。

表 5 算法模型分类指标表

Table 5 Classification metrics of algorithm model

分类器	XGBoost	RF
准确率	78.3%	74.2%
3 折交叉验证分数	73.9%	74.3%
平衡准确率	63.5%	57.9%
科恩卡帕系数	72.1%	66.6%
宏精准率	71.0%	66.0%
加权精准率	79.0%	74.0%
宏召回率	64.0%	58.0%
加权召回率	78.0%	74.0%
宏 F1 值	65.0%	59.0%
加权 F1 值	76.0%	72.0%

表 6 算法模型各证型分类指标表

Table 6 Classification metrics of each CM pattern in algorithm model

分类器	XGBoost			RF		
分类指标	精准率	召回率	F1 值	精准率	召回率	F1 值
0	89%	53%	67%	90%	60%	72%
1	86%	94%	90%	86%	94%	90%
2	64%	97%	77%	60%	97%	74%
3	0%	0%	0%	0%	0%	0%
4	89%	73%	80%	80%	55%	65%
5	71%	83%	77%	67%	67%	67%
6	100%	44%	62%	75%	33%	46%

在分类指标上，XGBoost 模型总的准确率、平衡准确率、科恩卡帕系数、宏精准率、加权精准率、宏召回率、加权召回率、宏 F1 值、加权 F1 值比 RF 模型高，但 3 折交叉验证分数比 RF 模型低。在各中医证型分类指标上，XGBoost 模型各证型的精准率、召回

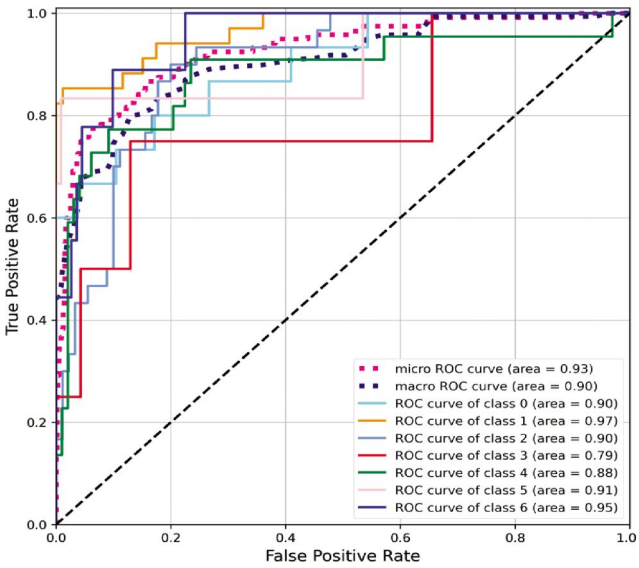

图 2 XGBoost 模型多分类 ROC 曲线图

Fig.2 Multi-classified ROC curves of XGBoost model

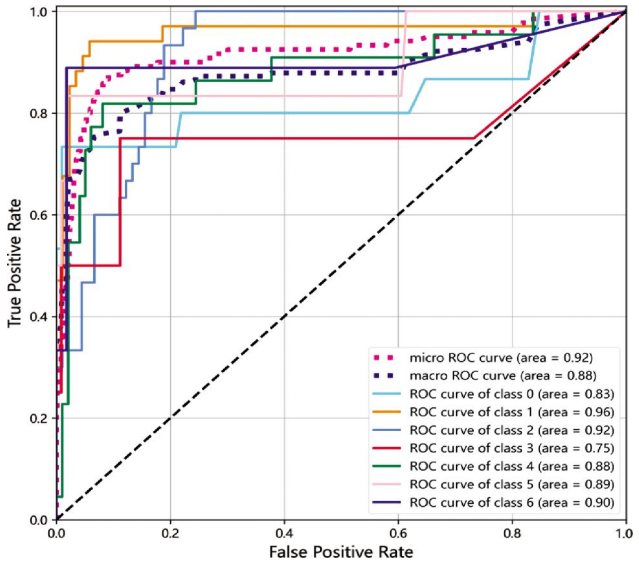

图 3 RF 模型多分类 ROC 曲线图

Fig.3 Multi-classified ROC curves of RF model

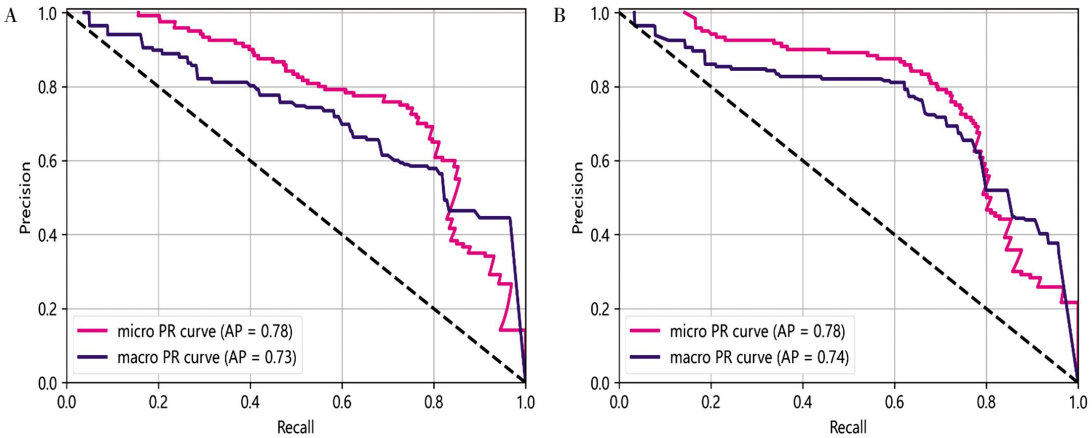

图 4 模型的 PR 曲线图

Fig.4 PR curves of the models

注：A. XGBoost 模型微、宏 PR 曲线图；B. RF 模型微、宏 PR 曲线图。

率、F1 值均比 RF 模型高。显然，XGBoost 模型各证型的分类指标整体优于 RF 模型。在性能曲线上，XGBoost 模型微、宏和各证型的 AUC 值比 RF 模型高，说明 XGBoost 模型微、宏和各证型的 AUC 值整体优于 RF 模型。XGBoost 模型微、宏 AP 值比 RF 模型低，说明在 PR 曲线上两者表现相当。

多项研究表明[19-22]，XGBoost 在分类预测中比贝叶斯网络、支持向量机、决策树和 RF 等算法准确率更高。本次结果显示，XGBoost 算法的分类指标和性能曲线评分也总体优于 RF 算法，可能是因为 XGBoost 算法与本数据集有更好的契合度。同时，该建模方法可为其他病证的证候客观化研究提供方法学指导，在证候研究中可能发挥重要价值。

参考文献

[1] TIAN J R, ZHANG D Y, YAO X, et al. Global epidemiology of systemic lupus erythematosus: A comprehensive systematic analysis and modelling study[J]. Annals of the Rheumatic Diseases, 2023, 82(3): 351-356.

[2] LI X B, HE Z Q, RU L, et al. Efficacy and safety of Qinghao Biejia Decoction in the treatment of systemic lupus erythematosus: A systematic review and meta-analysis[J]. Frontiers in Pharmacology, 2021, 12: 669269.

[3] WANG H Z, WANG B Z, HUANG J G, et al. Efficacy and safety of acupuncture therapy combined with conventional pharmacotherapy in the treatment of systemic lupus erythematosus: A systematic review and meta-analysis[J]. Medicine, 2023, 102(40): e35418.

[4] TIAN R, YUAN L, HUANG Y, et al. Perturbed autophagy intervenes systemic lupus erythematosus by active ingredients of traditional Chinese medicine[J]. Frontiers in Pharmacology, 2022, 13: 1053602.

[5] 夏淑洁, 杨朝阳, 周常恩, 等. 常见机器学习方法在中医诊断领域的应用述评[J].广州中医药大学学报, 2021, 38(4): 826-831.

[6] THABAH M M, SEKAR D, PRANOV R, et al. Neuromyelitis optica spectrum disorder and systemic lupus erythematosus[J]. Lupus, 2019, 28(14): 1722-1726.

[7] 郑筱萸. 中药新药临床研究指导原则: 试行[M]. 北京: 中国医药科技出版社, 2002: 111-115.

[8] 胡江帅. 基于贝叶斯网络技术对社区获得性肺炎的中医证型分析[D]. 昆明: 云南中医药大学, 2020.

[9] 姚乃礼. 中医症状鉴别学[M]. 北京: 人民卫生出版社, 2004: 1-300.

[10] 黎敬波, 马　力. 中医临床常见症状术语规范[M]. 北京: 中国医药科技出版社, 2005: 1-90.

[11] 李灿东. 中医诊断学[M]. 新世纪 4 版. 北京: 中国中医药出版社, 2016: 1-230.

[12] SHIN H. XGBoost regression of the most significant photoplethysmogram features for assessing vascular aging[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(7): 3354-3361.

[13] 董志阔. 系统性红斑狼疮中医体质与证素的相关性研究[D]. 天津: 天津中医药大学, 2023.

[14] ZHU J, NAUGHTON S, BOWMAN N, et al. Maternal antibody repertoire restriction modulates the development of lupus-like disease in BXSB offspring[J]. International Immunology, 2023, 35(2): 95-104.

[15] 宫爱民, 魏方志, 宋逸天. 系统性红斑狼疮中医证型及客观化研究进展[J]. 中医学, 2020, 9(2): 98-103.

[16] 王秀兰, 成佳黛, 卓进盛. 基于全国名中医林天东慢性咳嗽病案的中医证候与证素分布规律研究[J]. 中国民间疗法, 2021, 29(1): 4-5.

[17] 张冠壮, 黄宏敏, 许玉皎, 等. 海南地区中风病患者急性期中医证候的分布[J].世界中医药, 2017, 12(12): 3175-3178.

[18] 陈学武, 姜靖雯, 林海峰. 海南地区晚期非小细胞肺癌中医证候分布规律研究[J]. 海南医学, 2016, 27(4): 564-566.

[19] LI J L, LIU S R, HU Y D, et al. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: Retrospective cohort study[J]. Journal of Medical Internet Research, 2022, 24(8): e38082.

[20] WANG L Y, WANG X Y, CHEN A X, et al. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model[J]. Healthcare, 2020, 8(3): 247.

[21] HOU N Z, LI M Z, HE L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost[J]. Journal of Translational Medicine, 2020, 18(1): 462.

[22] SHIN H. XGBoost regression of the most significant photoplethysmogram features for assessing vascular aging[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(7): 3354-3361.

（本文编辑　周　旦）