DOI：10.13718/j.cnki.xdzk.2020.10.003

基于集成学习算法的慢性肾病早期筛查方法

姜玉苹[1]，　余　诚[2]，　林燕榕[1]，　斯海燕[1]，
刘　迪[1]，　朱　江[1]，　王　浩[1]，　陈　浩[1]

1. 肾泰网健康科技（南京）有限公司，南京 210023；2. 南京大学 软件新技术国家重点实验室，南京 210023

摘要：慢性肾病是严重危害人类健康的常见疾病，其发病率高，知晓率低. 基于集成学习算法的慢性肾病早期筛查方法能够提高肾病知晓率，有利于做到早发现早治疗. 搜集 2016 年到 2019 年多家医院的体检资料，选取 3 年内进展为慢性肾病的体检人员作为研究对象，并选取 3 年内没有进展为慢性肾病的体检人员作为对照组. 通过 5 折交叉验证，采用 python 3.7 进行随机森林与 XGBoost 算法模型的训练及测试，通过进展为慢性肾病结局的 $F1$ 值、真阳性和真阴性指标比较各模型对体检人员 3 年内是否进展为慢性肾病的预测效果. 随机森林算法模型预测效果为，真阳性率 0.950，真阴性率 0.969，$F1$ 值 0.957；XGBoost 算法模型预测效果为，真阳性率 0.966，真阴性率 0.955，$F1$ 值 0.958.

关　键　词：慢性肾病；早期筛查；集成学习；随机森林；XGBoost；交叉验证

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673－9868(2020)10－0017－08

慢性肾病具有患病率高、知晓率低、预后差和医疗费用高等特点，是继心脑血管疾病、糖尿病和恶性肿瘤之后，又一严重危害人类健康的疾病. 近年来慢性肾病患病率逐年上升，全球一般人群患病率已高达 14.3%[1]. 我国横断面流行病学研究显示[2]，18 岁以上人群慢性肾病患病率为 10.8%，据此估计我国现有成年慢性肾脏病患者 1.5 亿，但知晓率仅为 12.5%，该调查还发现经济快速发展的农村地区居民成为慢性肾脏病的高发人群. 随着我国人口老龄化、糖尿病和高血压等疾病的发病率逐年增高，慢性肾病发病率也呈现不断上升之势[3]. 由此可见对慢性肾病早期筛查的重要性. 随着人工智能技术的发展，越来越多的研究者将其应用到医疗卫生领域[4-9]. 人工神经网络、支持向量机、决策树等机器学习方法可以实现分类功能，并在疾病的风险预测方面得到应用. 而使用集成学习方法比单个机器学习方法构建的分类器性能表现更优[10-13]. 使用集成学习方法已经在各个领域实现图像识别、语义识别、疾病筛查、辐射源信号识别、天气预测等功能[14-17]. 基于集成学习算法的慢性肾病早期筛查方法在医疗领域具有重要价值.

1　研究方法

基于集成学习算法的慢性肾病早期筛查方法及其应用过程如图 1 所示，包含了医学数据的处理、模型的训练与测试及其应用部分.

收稿日期：2020－09－30
基金项目：江苏省重点研发(社会发展)项目(BE20191611)；南京市栖霞区发展和改革委员会第二批人工智能企业项目.
作者简介：姜玉苹(1991－)，女，硕士，主要从事大数据挖掘、医学图像处理研究.

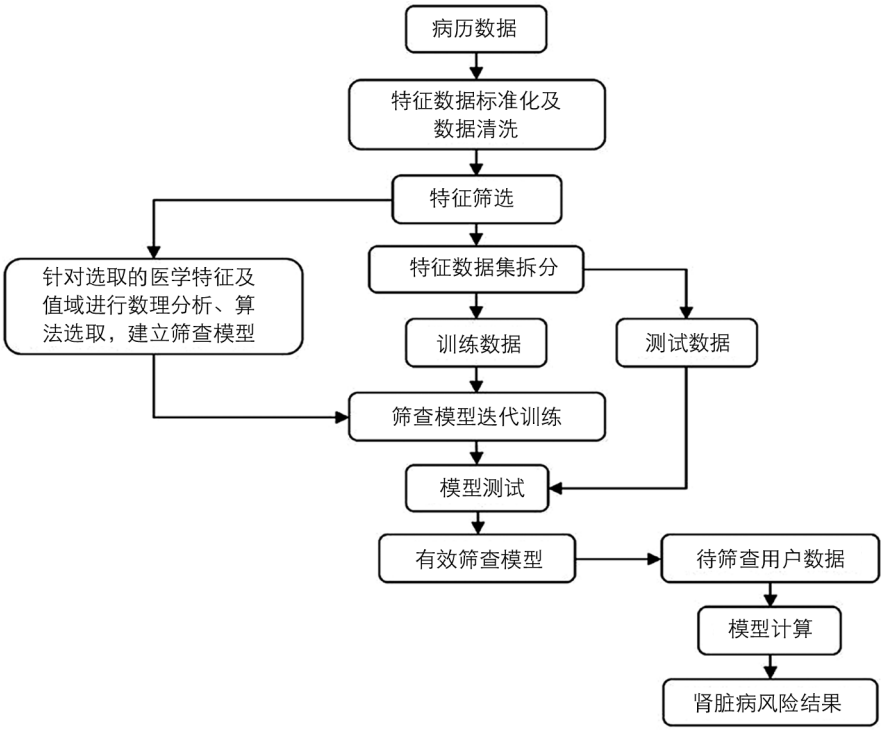


图1 肾病早期筛查模型训练、测试与应用流程

1.1 数据搜集与处理

本研究搜集2016年到2019年江苏省多个市的各个年龄段体检人员的体检资料，其内容包括身高、体质量、血压等基本信息和血常规、血生化、尿常规、肝肾功能、血脂等医学特征数据. 去除体检人员的身份证号码、电话号码、住址等敏感信息，排除了数据不完整和诊断不明确的体检资料，最终从中选取了3年内进展为慢性肾病的体检资料3 999例，没有进展为慢性肾病的体检资料4 536例.

数据处理包括标准化、特征筛选. ① 标准化：依据建立的医学实体标准库与单位的换算表，并对等级资料数据制定赋值规则. ② 特征筛选：将所有特征，按照规则分区间赋值，采用统计学上的卡方检验，根据检验结果删除没有统计学意义的特征. 根据肾脏病科医生提出的要求，将3年内进展为慢性肾病的数据标注为1，将3年内没有进展为慢性肾病的标注为0. 参与标注的人员共10人，均有医学相关专业背景，将其分为3组，每组标注人员都对全部数据进行标注. 最后，三组标注结果进行比对，结果有分歧的都会由肾脏病科医生给出最终的判断.

1.2 筛查模型构建

集成学习是一种组合多个基学习器共同完成预测任务的学习算法，与单个分类器相比具有更精确、过拟合风险低的优点，根据其组合思想可分为bagging和boosting两类[18-19]，本文分别选取经典模型随机森林和XGBoost进行构建[20-21]. 随机森林和XGBoost通常都是使用决策树作为基学习器进行集成. 决策树是最符合人类思考模式，最容易被理解和解释的模型之一，是一种树形结构，其中每个内部节点表示一个特征条件，每个分支代表对于这个特征条件的不同输出，每个叶节点代表结果相同的一类样本. 常见的决策树算法有C4.5,ID3和CART，能够有效地运行在大规模数据集上，处理具有高维特征的输入样本.

1.2.1 模型训练

将数据集按照一定比例分为训练集和测试集，训练集用来训练慢性肾病早期筛查模型，测试集用来测试模型的筛查性能. 使用随机森林算法训练过程中，会以有放回采样的方式为每个决策树提供一份不同的样本数据，每个决策树再计算结点上分支效果最佳的医学特征及阈值完成树的分裂构建，最终使得慢性肾病预测结果与相应患者标准诊断结果相符，从而得到慢性肾病早期筛查模型. 使用XGBoost算法训练过程中，所有决策树使用同一份数据集，但是要拟合的结果是上一轮模型输出与真实值之间的差距，同样对于每个决策树计算结点上分支效果最佳的医学特征及阈值完成树的分裂构建，最终使得慢性肾病预测结果与

相应患者标准诊断结果相符，从而得到慢性肾病早期筛查模型.

1.2.2　模型测试

使用超参数的网格搜索方法与 k 折交叉检验，进行多次模型训练进而优化模型. 网格搜索方法可以穷举参数的取值组合，再利用交叉验证评价优劣，最终获得最佳有效慢性肾病筛查模型.

交叉验证[22]方法可分为 2 种：留一交叉验证(leave-one-out cross-validation，LOO-CV)和 k 折交叉验证(k-CV). LOO-CV 通常是在数据缺乏的情况下使用，一般情况下都是使用 k-CV 方法. 将原始数据均分成 k 组，将每个子集数据分别做一次测试集，其余的子集数据作为训练集，这样会得到 k 个模型，k 个模型性能指标的平均值表示此 k-CV 下模型的性能. k-CV 可以有效地避免过拟合与欠拟合的发生，最后得到的结果较有说服性.

1.2.3　评估方法

评估筛查模型使用的度量指标：精确率(Precision，P)、真阳性率(True Positive Rate，TPR)、真阴性率(True Negative Rate，TNR)、$F1$ 值(F1-score). 精确率是指在慢性肾病筛查模型识别出的慢性肾病患者中被正确识别的占比，$P=TP/(TP+FP)$；真阳性率是指慢性肾病患者被正确识别的占比，$TPR=TP/(TP+FN)$；真阴性率是指非慢性肾病患者被正确识别的占比，$TNR=TN/(TN+FP)$；$F1$ 值为 P 和 TPR 的调和平均值，$F1=2*P*TPR/(P+TPR)$. $F1$ 值能够更好地体现模型的性能.

表 1　模型预测结果的混淆矩阵

样本分类	金标准为正样本	金标准为负样本
预测为正样本	TP	FP
预测为负样本	FN	TN

1.3　概率校准

随机森林、XGBoost 等分类器能够输出样本的后验概率为 $P(Y=c\mid X)$，其中 c 代表类别，X 代表样本特征. 但是模型输出的概率经常与样本属于类 c 的真实概率有偏差，本产品的应用场景为慢性肾病早期筛查，用户使用该软件能够检测自己患慢性肾病的风险概率，对模型的概率输出有较高要求，为消除偏差，需通过概率校准获取较准确的概率.

1.3.1　概率校准方法

概率校准即对分类预测的概率重新进行计算，再根据衡量标准评定校准结果的好坏. 为避免校准时引入偏差，校准模型需要用一个独立于训练集的测试集. 对于分类模型主要有 Platt scaling 和 Isotonic regression 2 种校准方式[23].

Platt scaling 是用 Logistics 回归对模型的输出值做拟合，该方法适用于样本量小的情况. 将模型输出值作为样本特征，与样本实际标签构成校准模型训练集，训练得到一维 LR 模型，最终输出的结果就是模型经过 Platt scaling 后的预测概率. 假设样本 i 的模型输出值为 f_i，经 Platt scaling 后预测概率为：

$$P_{(1|f_1)}=\frac{1}{1+\exp(Af_i+B)}$$

其中，参数 A，B 可训练得到.

Isotonic 回归又称保序回归，是一种非参回归模型，该方法直接对 reliability 图中的数据进行拟合. 假设预测值 f_i 与实际标签 y_i 有：

$$y_i=m(f_i)+\varepsilon_i$$

约束条件：m 为单调递增函数. 目标函数：

$$\hat{m}=\underset{z}{\operatorname{argmin}}\sum(y_i-z(f_i))^2$$

目标函数的一种求解方法是 PAV 算法(pair-adjacent violators algorithm)，主要思想是输入训练集(f_i，y_i)并使 f_i 从小到大排序，之后不断合并调整违反单调性的局部区间，直到全局满足单调性.

1.3.2　概率校准衡量标准

由于真实条件概率未知，数据只有类别标签，DeGroot & Fienberg[24]提出了通过 reliability 图进行可

视化模型校准的方式，它可以大致评估出当前模型的输出结果与真实结果有多大偏差. 绘制该图有以下几个步骤：

步骤一：将概率空间分为 n 箱；

步骤二：各样本按模型得到的预测概率值分别落入 n 箱中；例如将概率空间分为10箱，预测概率为0～0.1的样本落入第一个箱中，预测概率为0.1～0.2的样本落入第二个箱中，以此类推.

步骤三：计算出各箱的预测平均值和各箱样本中的实际正样本比例，作为图中各箱的点的坐标.

步骤四：最后把图中各点连接起来，其斜率越接近1就意味着模型输出结果估计越有效.

另外，模型输出结果的Brier分数也可作为校准的一个衡量标准. 在二分类的情况下，计算Brier分数的公式：

$$BS=\frac{1}{N}\sum_{t=1}^{N}(f_t-o_t)^2$$

其中，f_t 是预测的事件 t 的发生概率；o_t 是事件 t 的实际标签(发生为1，不发生为0)；N 是预测事件总数量. 因此Brier分数越接近0，代表模型预测结果越接近于真实标签.

2 实验结果

2.1 数据集构造

对搜集到的数据进行单位和数值等标准化处理后，对其进行卡方检验或 T 检验统计学分析. 表2展示了慢性肾病在各个特征不同取值的是否进展为慢性肾病结局的分布情况，及其统计学检验结果. 统计结果显示，进展为慢性肾病人群集中在老年人群，并且进展为慢性肾病的血压、血糖、血肌酐、尿素与尿蛋白水平均高于没有进展为慢性肾病的，而进展为慢性肾病的白蛋白与总蛋白水平明显低于没有进展为慢性肾病的. 根据卡方检验结果显示，每个特征针对3年内是否进展为慢性肾病都是具有统计学意义的($p<0.01$)，表明选取的特征与慢性肾病存在极其显著的相关关系，选取这些特征建立模型是合理的.

表2 病例统计

变量		没有进展为肾脏病 N/%	进展为肾脏病 N/%	p 值	变量		没有进展为肾脏病 N/%	进展为肾脏病 N/%	p 值
age	20～39	2 483(97.0)	77(3.0)	<0.001	收缩压	<90	24(48.0)	26(52.0)	<0.001
	40～64	1 577(72.4)	602(27.6)			90～138	4 463(53.7)	3 845(46.3)	
	65～79	240(9.4)	2 311(90.6)			>139	3(7.7)	36(92.3)	
	>80	236(19.0)	1 008(81.0)		舒张压	<60	344(94.0)	22(6.0)	<0.001
sex	男	2 815(59.6)	1 908(40.4)	<0.001		60～89	3 700(55.7)	2 940(44.3)	
	女	1 721(45.1)	2 091(54.9)			>90	445(39.2)	689(60.8)	
BMI	<24	2 316(65.9)	1 196(34.1)	<0.001	总蛋白	<60	1(25.0)	3(75.0)	<0.001
	24～27	1 781(44.0)	2 265(56.0)			60～79	3 601(72.4)	1 373(27.6)	
	>28	438(44.9)	537(55.1)			>80	524(87.0)	78(13.0)	
白蛋白	<35	6(21.4)	22(78.6)	<0.001	尿素	<2.9	122(41.9)	169(58.1)	<0.001
	35～50	3 225(65.1)	1 730(34.9)			2.9～7.4	4 228(57.7)	3 105(42.3)	
	>51	912(92.8)	71(7.2)			>7.5	182(23.6)	589(76.4)	
白细胞计数	<4	73(25.2)	217(74.8)	<0.001	空腹血糖	<3.9	33(31.4)	72(68.6)	<0.001
	49～	4 076(63.8)	2 314(36.2)			3.9～6.0	4 021(58.8)	2 823(41.2)	
	>10	369(73.9)	130(26.1)			>6.1	482(32.0)	1 024(68.0)	
总胆固醇	<3	69(46.3)	80(53.7)	<0.001	尿隐血	−	4 030(71.0)	1 648(29.0)	<0.001
	3～5.68	3 829(57.7)	2 810(42.3)			+−	162(48.5)	172(51.5)	
	>5.69	633(42.2)	866(57.8)			1+	177(51.6)	166(48.4)	

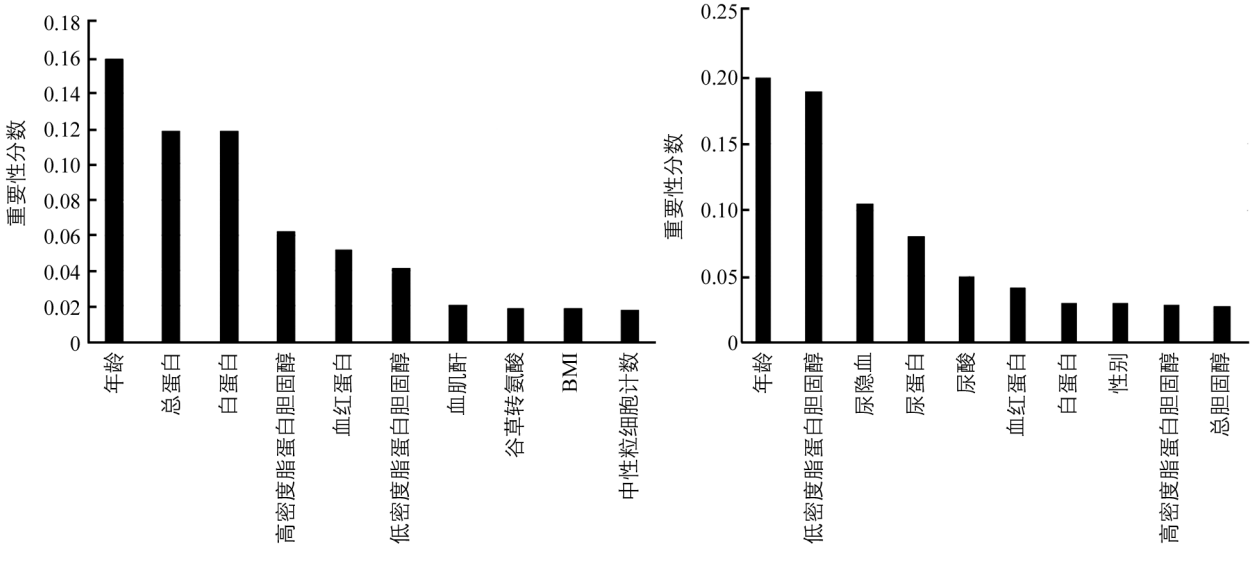
续表 2

变量		没有进展为肾脏病 N/%	进展为肾脏病 N/%	p 值
血肌酐	＜40	4 192(55.3)	3 389(44.7)	＜0.001
	40～49	198(47.7)	217(52.3)	
	50～100	20(25.3)	59(74.7)	
	＞101	126(27.4)	334(72.6)	
尿酸	＜149	23(29.9)	54(70.1)	＜0.001
	149～415	3 708(72.1)	1 434(27.9)	
	＞416	691(64.8)	376(35.2)	
	2＋	108(52.9)	96(47.1)	
	3＋	59(47.6)	65(52.4)	
尿蛋白	－	3 781(67.9)	1 784(32.1)	＜0.001
	＋－	525(76.8)	159(23.2)	
	1＋	172(59.9)	115(40.1)	
	2＋	47(39.2)	73(60.8)	
	3＋	10(38.5)	16(61.5)	

2.2　筛查模型训练

根据卡方检验分析结果，最终纳入训练模型的特征有性别、年龄、BMI、血压、血白细胞计数、血中性粒细胞计数、血红蛋白、血白蛋白、血总蛋白、血甘油三脂、血总胆固醇、血高密度脂蛋白胆固醇、血低密度脂蛋白胆固醇、血肌酐、血尿酸、血谷草转氨酶、血谷丙转氨酶、血尿素、空腹血糖、尿隐血、尿蛋白. 使用超参数网格搜索方法优化筛查模型，其中随机森林涉及到的主要参数有 n_estimators，max_depth，min_samples_leaf，min_samples_split，max_features，criterion，XGBoost 涉及到的主要参数有 learning_rate，n_estimators，min_child_weight，subsample，colsamole_bytree，gamma，reg_alpha，reg_lambda. 并使用 5 折交叉检验方法训练测试筛查模型，取 5 次的筛查模型测试结果的平均值作为最终测试评估结果.

采用随机森林和 XGBoost 算法模型训练时计算给出的每个特征重要性分数如图 2 所示，使用随机森林算法计算得出年龄、总蛋白、白蛋白对慢性肾病早期筛查的作用较大，使用 XGBoost 算法计算得出年龄、低密度脂蛋白胆固醇、尿隐血对慢性肾病早期筛查的作用较大.



(a) 随机森林　　(b) XGBoost算法模型

图 2　随机森林(左)与 XGBoost(右)算法模型的特征重要性评估

2.3　筛查模型测试

表 3 展示了 5 折交叉检验方法得到的基于集成学习算法的慢性肾病早期筛查方法的性能. 使用随机森林算法的慢性肾病早期筛查方法的精确率、真阳性率、真阴性率和 $F1$ 值分别为 0.964，0.950，0.969，0.957，使用 XGBoost 算法的分别为 0.950，0.966，0.955，0.958.

表 3　模型性能比较

评估指标	*RF*	*XGBoost*	评估指标	*RF*	*XGBoost*
精确率	0.964	0.950	真阴性率	0.969	0.955
真阳性率	0.950	0.966	$F1$ 值	0.957	0.958

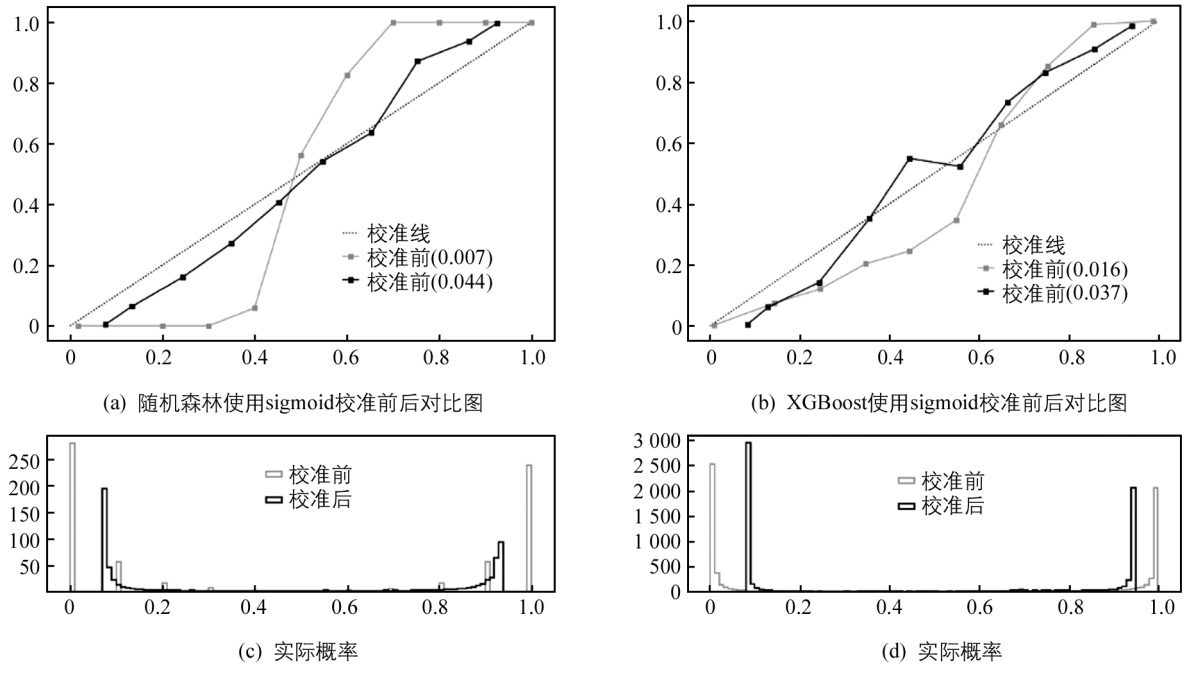
注：*RF*：随机森林算法模型；*XGBoost*：*XGBoost* 算法模型.

2.4 筛查模型概率校准结果

未校准的随机森林输出预测概率的 reliability 图和分布直方图，如图 3 左图所示. 未校准的随机森林 Brier 分数 0.007，模型预测结果集中于 0 和 1 这 2 个值，reliability 图呈明显的 S 型，说明模型分类性能好，但输出的概率的可靠性不佳. 选择 Platt scaling 进行概率校准，如图 3 左图所示. 校准后 Brier 分数为 0.044，reliability 图上的各点连接的曲线明显接近对角线，随机森林输出的概率估计准确性经校准后得到一定提升，这说明 Platt scaling 可以有效地降低随机森林的偏差.

未校准的 XGBoost 输出预测概率的 reliability 图和分布直方图如图 3 右图所示. 未校准的情况下，XGBoost 模型的 Brier 分数为 0.016，与随机森林一样，XGBoost 预测值也有很多落在 0，1 这 2 处，导致 Brier 分数很小，整个曲线比随机森林更接近对角线，概率可靠性略有提升. 同样选择 Platt scaling 进行概率校准，如图 3 右图所示. 校准后的 Brier 分数为 0.037，虽然上升了 0.021，但是 reliability 图上的曲线总体上更贴近对角线，概率校准在一定程度上提高了概率估计的有效性，且基本没有影响原模型的分类性能.

随机森林与 XGBoost 算法模型采用 Platt scaling 概率校准方法校准前后，对慢性肾病的早期筛查性能的比较见表 4. 校准后的模型性能存在一定程度的下降，但是均高于 0.94.



(a) 随机森林使用sigmoid校准前后对比图　(b) XGBoost使用sigmoid校准前后对比图

(c) 实际概率　(d) 实际概率

图 3 概率校准前后的 reliability 图和分布直方图

表 4 模型概率校准前后性能比较

评估指标	*RF*	*RF*＋*Platt*	*XGBoost*	*XGBoost*＋*Platt*
Brier	0.007	0.044	0.016	0.037
精确率	0.999	0.940	0.970	0.941
真阳性率	0.997	0.950	0.985	0.968
真阴性率	0.999	0.945	0.976	0.952
*F*1 值	0.998	0.945	0.978	0.955

注：*RF*：随机森林算法模型；*RF*＋*Platt*：随机森林算法模型采用 Platt scaling 概率校准方法校准；*XGBoost*：*XGBoost* 算法模型；*XGBoost* ＋*Platt*：*XGBoost* 算法模型采用 Platt scaling 概率校准方法校准.

3 结　论

本研究基于随机森林与 XGBoost 集成学习算法创建慢性肾病早期筛查方法，使用随机森林算法训练得到的筛查模型精确率、真阳性率、真阴性率和 $F1$ 值分别为 0.964，0.950，0.969，0.957，XGBoost 算法的分别为 0.950，0.966，0.955，0.958. 其中随机森林算法的精确率与真阴性率较高，XGBoost 算法的真阳性率与 $F1$ 值较高. 总体来讲，2 种集成学习算法筛查模型性能相当，可以根据不同的筛查需求来选择. 该慢性肾病早期筛查方法在应用过程中，2 个模型共同筛查得到的阳性结果就可以判定为阳性.

慢性肾病筛查最终得出的结果是患者发展为慢性肾病的风险概率值，而分类模型直接输出的分数值并不能直接视为风险预测的概率值，需要评估出当前模型的输出结果与真实结果的偏差是否在允许的范围内，必要的时候需要对其结果进行校准，因此选用概率校准方法解决这个问题. 本文使用 Platt scaling 概率校准方法校准后的模型性能存在一定程度的下降，但是均高于 0.94.

由于给出的数据并不知道患者患慢性肾病的真实概率值，无法直接判断原模型的输出是否为有效估计，一种简单而普适的方法即绘制 reliability 图，图线越接近对角线，说明模型的概率估计越有效，若超出预期范围，可以采用 Platt scaling 概率校准方法来降低原分类模型的偏差，使最终输出值更接近真实概率，经过概率校准处理后使原模型最终的输出是有效的估计值.

综上，基于随机森林、XGBoost 集成学习算法的慢性肾病早期筛查方法的预测效果均表现良好且稳定. 采用 Platt scaling 概率校准方法进行模型概率校准并没有过多的改变分类性能，只是提升了原模型对慢性肾病风险概率估计的可靠性，因此概率校准后输出的概率值更具临床参考价值. 基于集成学习算法的慢性肾病早期筛查方法可以应用于医院、体检中心、社区、保险公司及移动平台等辅助体检人员的慢性肾病早期筛查.

参考文献：

[1] ENE-IORDACHE B，PERICO N，BIKBOV B，et al. Chronic Kidney Disease and Cardiovascular Risk in Six Regions of the World (ISN-KDDC)：a Cross-Sectional Study [J]. Lancet Glob Health，2016，4(5)：e307-e319.

[2] ZHANG L，WANG F，WANG L，et al. Prevalence of Chronic Kidney Disease in China：a Cross-Sectional Survey [J]. Lancet，2012，379(9818)：815-822.

[3] 上海慢性肾脏病早发现及规范化诊治与示范项目专家组，高　翔，梅长林. 慢性肾脏病筛查诊断及防治指南 [J]. 中国实用内科杂志，2017，37(1)：28-34.

[4] 洪　烨. 基于机器学习算法的糖尿病预测模型研究 [D]. 哈尔滨：哈尔滨工业大学，2016.

[5] 周悦玲，蒋更如. IgA 肾病进展至终末期肾病临床预测的研究现状 [J]. 上海交通大学学报(医学版)，2016，36(2)：296-301.

[6] 刘迷迷，蔡永铭. 基于多层感知神经网络的糖尿病并发症预测研究 [J]. 软件，2018，39(10)：30-35.

[7] 郑晓燕. 基于机器学习的心血管疾病预测系统研究 [D]. 北京：北京交通大学，2018.

[8] 刘　璐. 基于机器学习的小于胎龄儿预测模型的研究 [D]. 北京：北京工业大学，2017.

[9] 周　超. 基于机器学习的感知信号分类与预测方法研究 [D]. 成都：电子科技大学，2018.

[10] 方育柯，傅　彦，周俊临. 基于集成学习的个性化推荐算法 [J]. 计算机工程与应用，2011，47(10)：1-4.

[11] 谭言丹，赵阳洋，赵光财. 基于 AdaBoost 特征选择和 XGBoost 的帕金森病诊断 [J]. 信息技术，2020，44(9)：124-128.

[12] 侯　勇，郑雪峰. 集成学习算法的研究与应用 [J]. 计算机工程与应用，2012，48(34)：17-22.

[13] 李　勇，刘战东，张海军. 不平衡数据的集成分类算法综述 [J]. 计算机应用研究，2014，31(5)：1287-1291.

[14] 李明峰，贾修一. 基于多分类器集成学习的中文反语识别技术 [J]. 计算机与数字工程，2018，46(9)：1790-1795.

[15] 刘　毅. 基于集成学习算法的冠心病早期筛查方法研究 [D]. 济南：山东大学，2018.

[16] 黄颖坤，金炜东，余志斌，吴昀璞. 基于深度学习和集成学习的辐射源信号识别 [J]. 系统工程与电子技术，2018，40(11)：2420-2425.

[17] 李俊磊. 多组合分类器在局部区域气温预测中的研究与应用[D]. 广州：广东工业大学，2014.
[18] FAWCETT T. An Introduction to ROC Analysis [J]. Pattern Recognition Letters，2006，27(8)：861-874.
[19] BREIMAN L. Bagging Predicators [J]. Machine Learning，1996，24(2)：123-140.
[20] BREIMAN L. Random Forests [J]. Machine Learning，2001，45(1)：5-32.
[21] CHEN T，GUESTRIN C. XGBoost：A Scalable Tree Boosting System [C] //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，San Francisco California USA. New York，NY，USA ACM，2016：785-794.
[22] 范永东. 模型选择中的交叉验证方法综述[D]. 太原：山西大学，2013.
[23] NICULESCUMIZIL A，CARUANA R. Predicting Good Probabilities with Supervised Learning [C] //International Conference on Machine Learning，ICML'05，August 7-11，2005. Bonn，Germany. New York，USA：ACM Press，2005：625-632.
[24] DEGROOT M H，FIENBERG S E. The Comparison and Evaluation of Forecasters [J]. Journal of the Royal Statistical Society：Series D (The Statistician)，1983，32(1-2)：12-22.

An Early Screening Method of Chronic Kidney Disease Based on Ensemble Learning Algorithm

JIANG Yu-ping[1]，YU Cheng[2]，LIN Yan-rong[1]，SI Hai-yan[1]，LIU Di[1]，ZHU Jiang[1]，WANG Hao[1]，CHEN Hao[1]

1. *ShenTaiWang Healthcare Technology Limited Company，Nanjing 210023，China*；
2. *National Key Laboratory for Novel Software Technology at Nanjing University，Nanjing University，Nanjing 210023，China*

Abstract：Chronic kidney disease，with its high incidence and low awareness，is a common disease that seriously endangers human health. The early screening method of chronic kidney disease based on the ensemble learning algorithm can improve the awareness rate of kidney disease and is conducive to early detection and early treatment. In a study reported herein，the medical examination data of many hospitals from 2016 to 2019 were collected，the examinees who had progressed to chronic kidney disease within three years were selected as the research subjects，and the examinees who had not progressed to chronic kidney disease within three years were taken as the control group. Through 5-fold cross-validation，python 3.7 was used to train and test the random forest and XGBoost algorithm models，and their predictive effect was compared based on the $F1$-score，and true positive and true negative indicators of the outcome of chronic kidney disease. The prediction effect of the random forest algorithm model was that the true positive rate was 0.950，the true negative rate was 0.969 and the $F1$-score was 0.957；while that of the XGBoost algorithm model was that the true positive rate was 0.966，the true negative rate was 0.955 and the $F1$-score was 0.958.

Key words：chronic kidney disease；early screening；ensemble learning；random forest；XGBoost；cross validation

责任编辑 胡 杨