DOI:10.3969/j.issn.1005-202X.2022.06.021

医学生物信息

糖尿病视网膜病变的风险揭示与关键因素分析

申思源，罗冬梅
安徽工业大学数理科学与工程学院，安徽 马鞍山 243002

【摘要】目的：通过构建组合模型对糖尿病并发视网膜病变（DR）的患病风险进行预测，为DR的预防和诊断提供参考。**方法：**基于3 000例糖尿病患者的生化检测数据，运用互信息作为评价标准筛选出与DR有关的特征因素，将其作为入模变量构建5种常见的模型，以准确率、精确率、召回率和AUC作为评价标准筛选出预测能力较优的3种模型，并运用Stacking方法构建组合模型。**结果：**通过互信息筛选出39个特征因素，发现随机森林模型、SVM模型以及Logistic回归模型这3种模型表现较优；构建的3种组合模型中，发现以SVM、Logistic为初级分类器，随机森林为次级分类器的组合模型预测能力最好，其AUC高达0.877。**结论：**组合模型相比单一模型具有更好的DR风险预测能力，更有助于DR的临床诊断。

【关键词】糖尿病视网膜病变；互信息；组合模型；Stacking方法；风险预测
【中图分类号】R318；R587.1 **【文献标志码】**A **【文章编号】**1005-202X（2022）06-0783-05

Risk disclosure and key factors analysis of diabetic retinopathy

SHEN Siyuan, LUO Dongmei
School of Mathematics and Physics, Anhui University of Technology, Ma'anshan 243002, China

Abstract: Objective To construct combination models for the risk prediction of diabetic retinopathy (DR), thereby providing a reference for the prevention and diagnosis of DR. **Methods** Based on the biochemical test data of 3 000 diabetic patients, the characteristic factors related to DR which were screened out by taking mutual information as the evaluation criterion were input as the modeling variables to construct 5 common models. According to the evaluation criteria of accuracy rate, precision rate, recall rate and AUC, 3 models with the superior prediction ability were screened out for model combination by Stacking method. **Results** A total of 39 characteristic factors were screened out through mutual information. The prediction performances of random forest model, SVM model and Logistic regression model were better. Among 3 combination models constructed, it was found that the combination model with SVM and Logistic as the primary classifier and random forest as the secondary classifier had the optimal prediction ability, and its AUC was as high as 0.877. **Conclusion** The combination model is superior to the single model in predicting DR risk, and is helpful for the clinical diagnosis of DR.

Keywords: diabetic retinopathy; mutual information; combination model; Stacking method; risk prediction

前 言

糖尿病并发视网膜病变（Diabetic Retinopathy, DR）是糖尿病并发症中最常见的微血管并发症之一，属于糖尿病的衰弱并发症，患者的糖尿病病史越长，其发生DR的概率就越高[1]。DR患者的视网膜微血管系统易被破坏，会出现毛细血管基底膜增厚导致肿胀变形、血-视网膜屏障破坏，周皮细胞和内皮细胞死亡等症状[2]。若病情进一步发展，新生血管可致使视网膜微血管系统扭曲，导致视网膜脱离，甚至失明[3]。近些年来，全球的DR患者正逐年增加，预计到2030年全球将有3亿的DR患者[4]。

目前对于DR的临床诊断有眼底照相和荧光素钠眼底血管造影，荧光素钠眼底血管造影检查通过对糖尿病患者视网膜循环情况、血-视网膜屏障状态等进行评价，从而判断患者是否患有视网膜病[5]。除此之外，机器学习和人工智能相关方法在DR诊断中的应用也越来越普遍。如Gunasekeran等[6]利用人工智能建立DR病变个体风险模型，并用其对患者进行风险分层；Schneck等[7]建立基于多焦视网膜电流图

【收稿日期】2021-12-02
【基金项目】国家级创新创业训练项目（201910360121）；安徽省自然科学基金（1808085MG220）；安徽省教学研究项目（2020jyxm0238）
【作者简介】申思源，研究方向：机器学习，E-mail: 2391479819@qq.com
【通信作者】罗冬梅，博士，讲师，研究方向：数据科学，E-mail: luodmahut@126.com

隐式时间延迟的多变量模型，并用其预测非增殖型DR局部斑块的发展。另外，Somasundaram等[8]设计了一种Bagging集成分类器ML-BEC，较好地实现早期DR病变的筛选；而Zhang等[9]则利用机器学习算法对60种血浆细胞进行因子分析，不仅得到与DR病变强相关的3种因子，还构建了具有很好预测能力的随机森林模型。

不同于以往学者利用机器学习方法进行特征筛选[10-14]，本研究采用互信息作为工具衡量各个特征因素与糖尿病患者是否发生视网膜病变（label变量）之间的依赖性，并根据依赖性筛选出DR的关键因素，然后将它们作为入模变量，构建5种常见的集成学习模型，最后将预测能力较强的3种模型通过Stacking方法构建组合模型。相校单一预测模型，组合模型的预测能力更强。

1 对象与方法

1.1 数据来源及介绍

本研究所用数据来源于国家人口与健康科学数据共享临床医学科学数据中心（http://www.ncmi.cn）（301医院）提供的DR数据集。数据集包含了3 000名糖尿病患者的87项生化检测数据，如血尿素、脂蛋白、尿肌酐、糖化血红蛋白等，还包含患者的其他患病情况，例如高血压、高脂血、肾病、肺部肿瘤、冠心病等。

表1展示了3 000例患者的年龄分布，DR患者主要集中在40~79岁。男性患者共有1 874人，占比62.5%，其中约有49.8%的患者患有视网膜病变；女性患者共有1 126人，占比37.5%，其中约有50.3%的患者患有视网膜病变，说明男女性糖尿病患者患有视网膜病变的几率相差不大。

1.2 方法

首先对数据进行异常值、缺失值检查，对异常值进行删除操作，对缺失值采用K-最近邻算法[15]进行填补；然后计算每个特征与label变量之间的互信息，绘制条形图并筛选出与label变量具有强依赖性的特征为关键因素；其次将筛选出的关键因素作为入模变量，构建5种集成学习模型，并从中筛选出预测能力排名前3的模型；最后利用Stacking方法建立3种较优单一模型的组合模型，并利用准确率、精确度、召回率、AUC值对组合模型进行综合评价。

表1 3 000例患者年龄分布

Table 1 Age distribution of 3 000 patients

年龄区间/岁	总数	DR患者数	DR患者占比/%
10~19	1	1	100
20~29	31	19	61
30~39	125	73	58
40~49	494	273	55
50~59	992	514	52
60~69	890	457	51
70~79	337	134	40
80~89	81	28	35
90~99	1	1	100

1.2.1 数据预处理 通过对数据集的检查，发现数据集中含有大量缺失数据，不含有异常值。为增加模型的稳定性，首先删除缺失数据超过66.6%的特征，删除后剩余71个特征；然后利用K-最近邻算法[15]对剩余特征的缺失数据进行插补。K-最近邻算法当K的取值选择合适时，在训练时就对异常点不敏感，并且它不是显式的训练，训练时间很短，适合大量数据插补。

1.2.2 变量筛选 互信息是信息论中的一个重要的信息度量，度量的是一个随机变量包含另一个随机变量的信息，可以表明随机变量之间的相互依赖性，两变量依赖越强，二者之间的互信息越大[16]。其计算公式为：

$$I(\xi;\eta)=\sum_{\xi,\eta}P(\xi,\eta)\times\log\frac{P(\xi,\eta)}{P(\xi)\times P(\eta)} \tag{1}$$

其中，ξ、η为两个随机变量，其联合分布为$P(\xi,\eta)$，边缘分布分别为$P(\xi)$、$P(\eta)$；$I(\xi;\eta)$是信息η（信宿收到）出现后提供的有关信息ξ（信源发出）的信息量，能够反映η对ξ的依赖性大小。

已有研究表明互信息可用于各个领域的特征选择且效果良好，对后续预测模型的建立、分类有重要帮助。如Wang等[17]选择与金属氧化物化学性质的特性具有最大互信息的特征集来对不同的化学物质进行分类；Samuel等[18]利用基于互信息的特征选择方法筛选出与中期电力负荷预测相关的特征，构建一个高精度的中期电力负荷预测模型；Rish等[19]将基于互信息的转导特征选择方法应用于遗传性状预测，取得优于其它特征选择方法的结果。

基于互信息的强大特征选择能力，本研究利用RStudio软件中的Fselector包计算预处理之后，计算71个因素与label变量之间的互信息，绘制条形图，最终得到与DR有关的39个关键因素。

1.2.3 单一模型构建及选择 从整理好的包含39个特征的3 000例病患数据中随机抽取70%作为训练集、30%作为测试集，分别利用随机森林模型[20]、梯度提升决策树（Gradient Boosting Decision Tree, GBDT）模型[21]、Logistic回归模型[22]、XgBoost模型[23]以及支持向量机（Support Vector Machine, SVM）模型[24]

对数据进行训练验证，并以准确率、精确度、召回率、AUC值为评价标准选择出预测能力排名前3的模型。

1.2.4 Stacking方法构建组合模型　Stacking方法是通过增加算法的多样性泛化误差以提高模型的预测能力[25]。Stacking方法的基本思想是：选取若干个模型作为初级分类器，利用这些分类器对原始数据进行训练测试，得到一系列新的预测值；然后将这些新的预测值作为新的特征加入到原始数据中，这样在训练时，数据集中又增加了与label变量具有强依赖性的信息；最后利用次级分类器对新生成的数据进行训练，得到最终的模型。在初级分类器训练数据时采用的是5折交叉检验，该方法将数据分成5份，每次取出一份作为测试集，其余作为训练集[15]。这种交叉训练方法可以避免模型过拟合，增强模型的稳定性。本研究首先构建5种单一机器学习模型并进行筛选，然后选取其中最优3种模型通过Stacking方法构建组合模型。

2 结　果

2.1 互信息筛选危险因素

本研究首先计算出71个因素与label变量之间的互信息值，其中与label变量有依赖性关系的有39个，称为关键因素，剩余的特征因素由于与label变量没有依赖性，不考虑作为入模变量。为更充分地显示特征因素与label变量之间的依赖性关系，绘制了71个特征因素和label变量的互信息条形图（横坐标为对应的特征因素，纵坐标为各特征因素与label变量的互信息值），如图1所示。

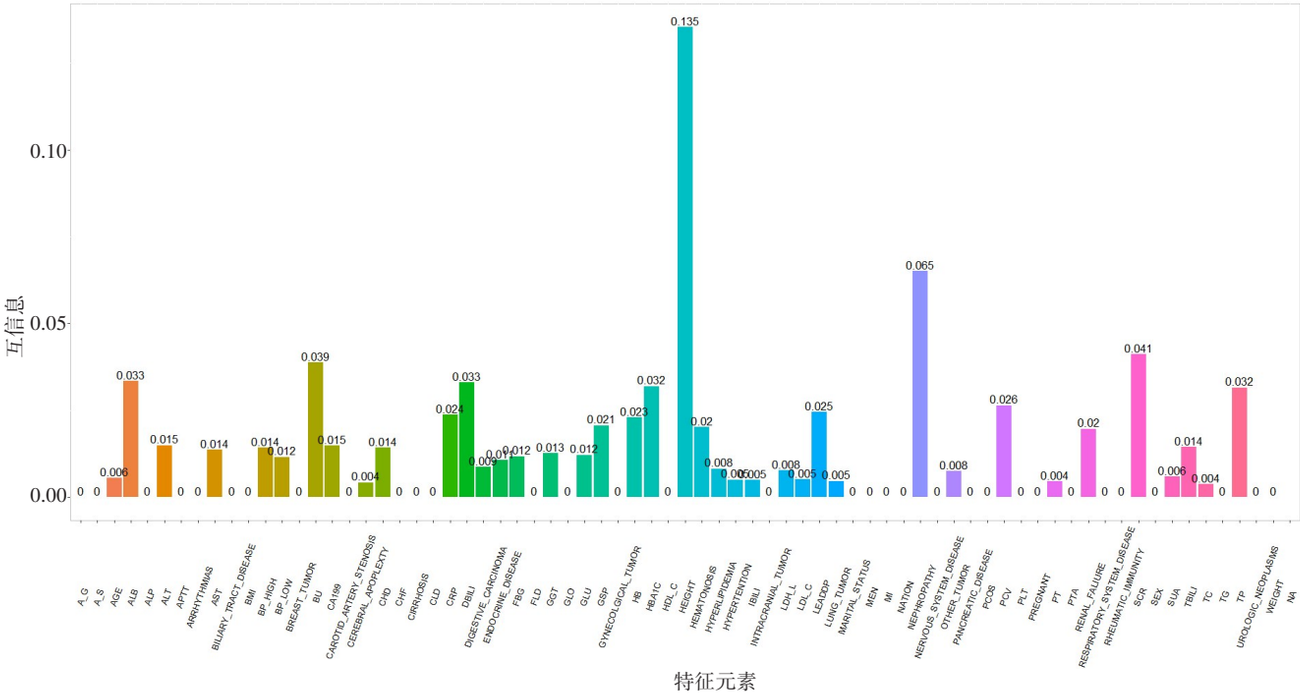


图1　特征因素与label变量的互信息

Figure 1 Mutual information between characteristic factors and label variables

共找到39种与DR相关的关键因素。其中HEIGHT（身高）、NEPHROPATHY（肾病）、SCR（血肌酐）、BU（血尿素）、ALB（血清白蛋白）、DBILI（直接胆红素）、TP（总蛋白）、HBA1C（糖化血红蛋白）、PCV（红细胞积压）、LEADDP（下肢动脉病变）、CRP（C反应蛋白）、HB（血红蛋白）与label变量具有较强的依赖性，这与曹文哲等[26]建模得到的危险因素相符合，说明互信息方法能有效筛选危险因素。

2.2 单一模型建模分析

利用R软件训练数据并构建5种模型，并用测试集检验模型，得到5种模型的准确率、精确度、召回率以及AUC，具体结果见表2。其中AUC是根据混淆矩阵计算得到特异度（Specificity）和召回率（Recall）绘制的ROC曲线下面积；准确度（Accuracy）、精确度（Precision）、召回率（Recall）、特异度（Specificity）的计算公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (5)$$

其中，TP表示真阳性的数量；TN表示真阴性的数量；FP表示假阳性的数量；FN表示假阴性的数量。

表2 5种模型的各项指标

Table 2 Indicators of 5 models

模型	准确率	精确度	召回率	AUC
随机森林	0.812	0.798	0.842	0.786
GBDT	0.761	0.848	0.623	0.759
Logistic回归	0.769	0.784	0.758	0.764
XgBoost	0.760	0.760	0.760	0.760
SVM	0.784	0.788	0.763	0.779

由表2可知，随机森林模型、Logistic回归模型和SVM模型具有较高的AUC值，分别为0.786、0.764、0.779。由于AUC主要用于综合评价模型的预测性能，鉴于以上3种模型的高AUC值，且其准确率、精确度、召回率也都处于较高水平，因此本研究选择随机森林模型、Logistic回归模型和SVM模型作为基础来构建组合模型。

2.3 建立组合模型

选取的3种单一模型（随机森林模型、Logistic回归模型和SVM模型）可以有3种组合方式来构建组合模型（表3）。Stacking方法将模型进行融合后，可以发挥3种算法的长处，并避免单一模型的短处，能够实现各种算法的取长补短，提升模型的预测能力。

表3 Stacking方法对模型组合结果

Table 3 Model combination by Stacking method

初级分类器	次级分类器	组合模型
Logistic回归和SVM	随机森林	1
SVM和随机森林	Logistic回归	2
Logistic回归和随机森林	SVM	3

图2展示了构建组合模型1的流程图。按照流程图中的步骤分别构建以上3种组合模型，利用构建的模型对测试集进行预测，得到混淆矩阵，然后根据混淆矩阵计算出3种组合模型的准确率、精确度、召回率（表4）。

由表4可以看出，组合模型1的准确率、召回率在3种组合模型的评价指标中是最高的，组合模型3次之，组合模型2最低；而精确度则组合模型2最高，组合模型3次之，组合模型1最低。采取第三级综合评价指标F_1-score对3种组合模型进行进一步评价，

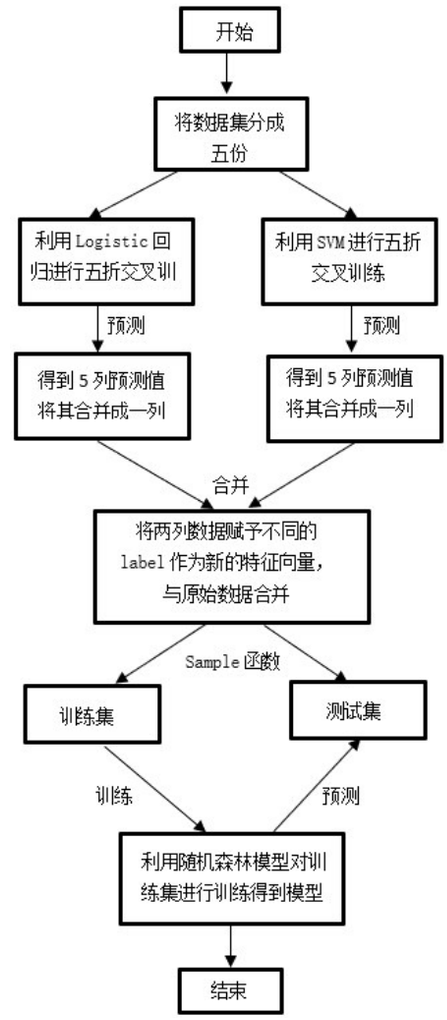


图2 组合模型构建流程图

Figure 2 Flowchart of combination model construction

表4 3种模型的相应指标

Table 4 Corresponding indicators of 3 combination models

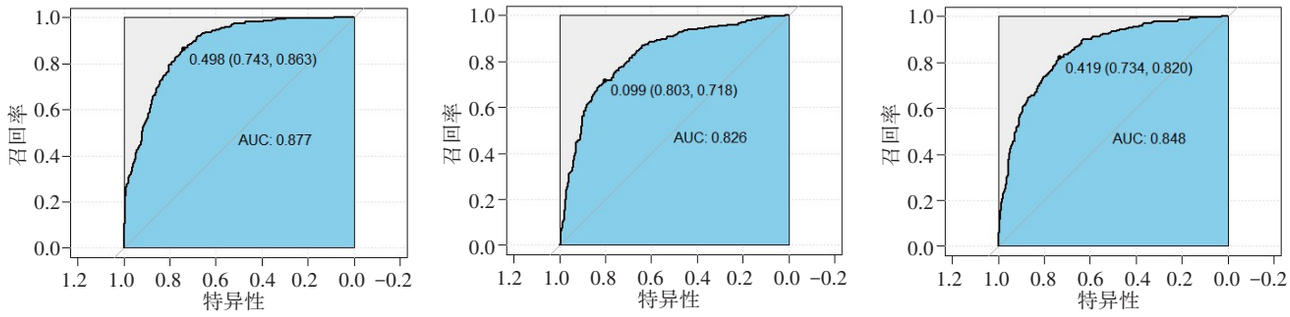
组合模型	准确率	精确度	召回率
1	0.803	0.758	0.864
2	0.753	0.820	0.675
3	0.757	0.791	0.724

其计算公式如下：

$$F_1\text{-score}=\frac{2\times P\times R}{P+R} \quad (6)$$

其中，P为精确度，R为召回率。F_1-score可以进一步评估模型的准确率。根据表4结果进行计算，组合模型1、2、3的F_1-score分别为0.807、0.774、0.753，显然组合模型1的F_1-score最高，说明了组合模型1的预测性能最优。

根据3种组合模型的ROC曲线，计算所对应的AUC值（图3）。组合模型1的AUC值最高，组合模型3次之，组合模型2最低，且都达到了80%以上，均优于单一模型。



a：组合模型1　　b：组合模型2　　c：组合模型3

图3　3种组合模型的ROC曲线

Figure 3 ROC curves of 3 combination models

综合来看，构建的3种组合模型中，组合模型1的预测能力最优，即以SVM模型和Logistic回归模型为初级分类器，随机森林模型为次级分类器构建的模型预测能力最强。可以利用该模型对糖尿病患者是否患视网膜病变进行风险揭示。

3 结 论

当前关于DR的研究可以分为两类：(1)根据眼底相机或多焦视网膜电流图等收集到的图像对DR进行智能诊断以及对患有DR的患者进行DR发展状况评估；(2)对DR患者的生化数据进行特征选择，根据选择出的关键因素建立预测分类模型。但总体而言，基于图像对DR预测模型的研究(图像处理计算量大，对计算设备要求高)更多，而且其预测能力也很突出；而基于关键因素建立预测模型的研究(计算量相对较少，计算时间较短，对计算设备要求不高)相对较少且预测能力一般。本研究通过Stacking方法构建多个单一模型的组合模型，不仅符合基于关键因素构建预测模型的优点(计算负担小)，其预测精度也得到了极大的提升(AUC达0.8以上)。更值得一提的是，本研究首次采用互信息对与DR有关的关键因素进行筛选，且效果良好，筛选出的与label变量具有较强的依赖性的因素，与其它研究通过模型筛选出的危险因素相一致[27]，操作简单。本研究构建的组合模型1科学合理，且能以较高的准确率预测糖尿病患者是否患有视网膜病变，有助于DR患者的筛检和预防，具有极大的临床应用价值。

【参考文献】

[1] 秦萍瑛. 糖尿病视网膜病变[J]. 世界最新医学信息文摘, 2018, 18(27): 33-35.
Qin PY. Diabetic retinopathy[J]. World Latest Medicine Information, 2018, 18(27): 33-35.

[2] Kowluru RA. Diabetic retinopathy, metabolic memory and epigenetic modifications[J]. Vision Res, 2017, 139: 30-38.

[3] Calderon GD, Juarez OH, Hernandez GE, et al. Oxidative stress and diabetic retinopathy: development and treatment[J]. Eye, 2017, 31(8): 1-6.

[4] Nakajima M, Cooney MJ, Tu AH, et al. Normalization of retinal vascular permeability in experimental diabetes with genistein[J]. Invest Ophthamol Vis Sci, 2001, 42(9): 2110-2114.

[5] 张睿, 杨泽敏. 2型糖尿病视网膜病变患病情况及其危险因素[J]. 慢性病学杂志, 2021, 22(7): 989-992.
Zhang R, Yang ZM. Prevalence and risk factors of type 2 diabetic retinopathy[J]. Chronic Pathematdogy Journal, 2021, 22(7): 989-992.

[6] Gunasekeran DV, Ting DS, Tan GS, et al. Artificial intelligence for diabetic retinopathy screening, prediction and management[J]. Curr Opin Ophthalmol, 2020, 31(5): 357-365.

[7] Schneck ME, Bearse MA, Adams AJ, et al. A multifocal electroretinogram model predicting the development of diabetic retinopathy[J]. Prog Retin Eye Res, 2006, 25(5): 425-448.

[8] Somasundaram SK, Alli P. A machine learning ensemble classifier for early prediction of diabetic retinopathy[J]. J Med Syst, 2017, 41(12): 201-210.

[9] Zhang N, Cao B, Zhang Y, et al. Plasma cytokines for predicting diabetic retinopathy among type 2 diabetic patients *via* machine learning algorithms[J]. Aging, 2020, 13(2): 1972-1988.

[10] Sun X, Guo S. Association between diabetic retinopathy and interleukin-related gene polymorphisms: a machine learning aided meta-analysis[J]. Ophthalmic Genet, 2020, 41: 216-222.

[11] Ogunyemi O, Kermah D. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records[J]. AMIA Annu Symp Proc, 2015, 5: 983-990.

[12] Basu S, Johnson KT, Berkowitz SA. Use of machine learning approaches in clinical epidemiological research of diabetes[J]. Curr Diab Rep, 2020, 20: 80.

[13] Raju M, Pagidimarri V, Barreto R, et al. Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy[J]. Stud Health Technol Inform, 2017, 245: 559-563.

[14] Wang J, Bai Y, Xia B. Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning[J]. IEEE J Biomed Health Inform, 2020, 24: 3397-3407.

[15] Florimbi G, Fabelo H, Torti E, et al. Accelerating the K-nearest neighbors filtering algorithm to optimize the real-time classification of human brain tumor in hyperspectral images[J]. Sensors (Basel), 2018, 18: 2314.

[16] Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Trans Neural Netw Learn Syst, 1994, 5 (9): 537-550.

[17] Wang XR, Lizier JT, Nowotny T, et al. Feature selection for chemical sensor arrays using mutual information[J]. PLoS One, 2014, 9(3): 841-848.

[18] Samuel O, Alzahrani FA, Khan RJ, et al. Towards modified entropy mutual information feature selection to forecast medium-term load using a deep learning model in smart homes[J]. Entropy (Basel), 2020, 22(1): 68-76.

[19] Rish I, He D, Haws D, et al. MINT: mutual information based transductive feature selection for genetic trait prediction[J]. IEEE/ACM Trans Comput Biol Bioinform, 2016, 13(3): 578-583.

[20] Breiman L. Random forests[J]. Mach Learn, 2001, 45: 5-32.

[21] Cai WY, Guo JH, Zhang MY, et al. GBDT-based fall detection with comprehensive data from posture sensor and human skeleton extraction [J]. J Healthc Eng, 2020, 2020: 8887340.

[22] Bonte C, Vercauteren F. Privacy-preserving logistic regression training [J]. BMC Med Genomics, 2018, 11: 86.

[23] Li W, Yin Y, Quan X, et al. Gene expression value prediction based on XGBoost algorithm[J]. Front Genet, 2019, 10: 1077.

[24] Cao J, Wang M, Li Y, et al. Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment[J]. PLoS One, 2019, 14: e0215136.

[25] Liao Y, Peng Y, Shi S, et al. Early box office prediction in China's film market based on a stacking fusion model[J]. Ann Oper Res, 2020, 25 (6):1-18.

[26] 曹文哲, 应俊, 陈广飞, 等. 基于Logistic回归和随机森林算法的2型糖尿病并发视网膜病变风险预测及对比研究[J]. 中国医疗设备, 2016, 31(3): 33-38.
Cao WZ, Ying J, Chen GF, et al. Risk prediction and comparitive research of type 2 diabetes mellitus complicated with retinopathy based on Logistic regression and random forest algorithm[J]. China Medical Devices, 2016, 31(3): 33-38.

[27] 朱文广, 李映雪, 杨为群, 等. 基于K-折交叉验证和Stacking融合的短期负荷预测[J]. 电力科学与技术学报, 2021, 36(1): 87-95.
Zhu WG, Li YX, Yang WQ, et al. Short-term load forecasting based on the K-fold cross-validation and stacking ensemble[J]. Journal of Electric Power Science and Technology, 2021, 36(1): 87-95.

（编辑：谭斯允）