doi:10.3969/j.issn.1001-4616.2019.02.007

基于随机森林算法的吸毒人员甄别模型研究

顾海艳[1],王　权[2]

(1.江苏警官学院计算机信息与网络安全系,江苏 南京 210031)
(2.深圳前海全民健康医疗技术公司,广东 深圳 518000)

[**摘要**]　基于数据挖掘技术,利用脉搏波数据构建模型进行吸毒人员的甄别技术,是一项新技术研究. 对采集的脉搏波数据,在数据预处理的基础上,运用随机森林算法构建吸毒人员甄别模型,该模型准确率虽然较高,但查全率、F1 值均较低. 为此提出了改进的随机森林算法,主要包括 3 种改进策略:采用划分多组训练集和测试集进行交叉验证,运用下采样方案来平衡样本分布,选用多评判指标选定模型构建参数. 通过测试,根据准确率、查准率、查全率、F1 值等多项指标的比较,发现改进的随机森林判别模型性能得到明显提升.

[**关键词**]　数据预处理,随机森林,判别模型,查全率,F1 值
[**中图分类号**]TP183　[**文献标志码**]A　[**文章编号**]1001-4616(2019)02-0044-06

Research on Drug Addicts Screening Model Based on Random Forest Algorithm

Gu Haiyan[1], Wang Quan[2]

(1.Department of Computer Information and Cyber Security, Jiangsu Police Institute, Nanjing 210031, China)
(2.Shenzhen Qianhai General Health Technology Co., Ltd, Shenzhen 518000, China)

Abstract: Based on data mining technology, using pulse wave data to construct screening model to detect drug addicts is a new technology research. After the pre-processing of pulse wave data, the original Random Forest classification model is initially established with high accuracy, but with a relatively low recall rate and F1 score. To resolve this issue, an improved classification model is henceforth proposed. The improved model mainly involves three improvement strategies: firstly, perform cross-validation by dividing multiple training sets and test sets to obtain generalization errors; Secondly, balance the sample distribution using down-sampling techniques; And finally, select model parameters based on multi-criteria analysis. According to the evaluation results of accuracy, precision, recall rates, and F1 scores, the performance of the improved Random Forest classification model has been significantly improved.

Key words: data preprocessing, Random Forest, classification model, recall rate, F1 score

目前,国内外对吸毒者的甄别主要依靠血液、尿液、唾液的化学检测. 尽管使用这些体液的测试方法获得了较高精度,但是在实际操作过程中均存在一定问题. 血液检测需要专业检测仪器,设备较为昂贵、不便携带;尿液检测方便,但是需要嫌疑人配合,代谢周期之后(通常为 3 天至 1 周)检测不出曾有毒品摄入[1];唾液检测最方便,但唾液易受污染、准确度不易保障,另外唾液检测板目前较贵,大量普及应用有困难. 因此,需要研究新的检测方法以满足现代警务工作对吸毒者快速筛查的要求,一个可行的解决办法是通过收集和分析吸毒人员的脉搏波,研究、开发基于移动终端的检测系统来实现吸毒人员的快速甄别.

脉搏波检测作为一种非侵入性检测方法,在医学和临床检验中应用已有多年[2-3]. 另外,自 2001 年以来,关于吸毒者脉搏特征的相关研究也在不断出现. 随着大数据技术的发展,出现了利用脉搏波构建吸毒人员甄别的模型相关研究[4]. 吸毒嫌疑人的甄别是一个典型的二元分类问题,本文以吸毒者和非吸毒者的年龄、性别和脉搏波数据为数据集,研究基于随机森林算法构建甄别模型的方法,为实现吸毒者的快速检测提供技术方案.

收稿日期:2019-01-16.
基金项目:江苏省“十三五”高等学校重点建设学科建设专项基金(2016-0838)、江苏省公安厅重点科研项目(2017KX033Z).
通讯联系人:顾海艳,副教授,研究方向:数据挖掘,信息安全. E-mail:ghy7388@126.com

1 数据预处理

在采集300例吸毒人员和1 293例正常人员的脉搏波数据的基础上,对其进行特征提取,选择脉搏波上升斜率、脉搏波自心脏经主动脉到股动脉分支的传导时间、脉搏波主峰和反射点高度差等9项脉搏波特征数据以及年龄、性别共11项数据,分别以x_1、x_2、x_3、x_4、x_5、x_6、x_7、x_8、x_9、x_{10}、x_{11}表示(其中变量x_{11}为因子型,只有0、1个值,其他10项都是数值型变量). 对10项数值型特征变量进行多重共线性检查、变量偏度校正、标准化处理3项数据预处理工作.

1.1 多重共线性检查

多重共线性是指数据模型中的特征变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确[5]. 采用R软件中的cor( )函数对数据集中10项数值型特征变量进行多重共线性检验. 该函数采用的是皮尔森相关系数检验的方法,检验结果如表1所示.

表1 特征变量多重共线性检验表

Table 1 The result of multicollinearity test

	x_1	x_2	x_3	x_4	x_5	x_6	X_7	x_8	x_9	x_{10}
x_1	1.00	0.24	0.67	−0.03	0.03	0.00	0.33	−0.29	0.26	0.06
x_2	0.24	1.00	0.32	−0.29	0.02	−0.42	0.14	−0.20	0.29	0.08
x_3	0.67	0.32	1.00	−0.11	0.29	−0.13	−0.12	−0.34	0.32	0.34
x_4	−0.03	−0.29	−0.11	1.00	0.70	0.69	−0.33	−0.13	−0.19	0.01
x_5	0.03	0.02	0.29	0.70	1.00	0.38	−0.51	−0.21	−0.04	0.24
x_6	0.00	−0.42	−0.13	0.69	0.38	1.00	−0.42	−0.20	−0.21	0.06
x_7	0.33	0.14	−0.12	−0.33	−0.51	−0.42	1.00	0.16	0.06	−0.35
x_8	−0.29	−0.20	−0.34	−0.13	−0.21	−0.20	0.16	1.00	−0.22	−0.24
x_9	0.26	0.29	0.32	−0.19	−0.04	−0.21	0.06	−0.22	1.00	0.52
x_{10}	0.06	0.08	0.34	0.01	0.24	0.06	−0.35	−0.24	0.52	1.00

表1结果表明,各特征变量之间的相关系数的绝对值均小于0.7,表明特征变量之间的线性相关性较弱,可将这些特征变量代入随机森林模型.

1.2 变量偏度校正

偏度是统计数据分布偏斜方向和程度的度量,是统计数据分布非对称程度的数字特征[6]. 建模中对于连续的响应变量不满足正态分布的情况要进行数据变换,通过变换,可在一定程度上减小不可观测的误差和预测变量的相关性,明显地改善数据的正态性、对称性和方差相等性. 由于特征变量存在偏度,对上述特征变量采用Box-Cox变换进行偏度变换,经过变换的特征变量比原特征变量更加接近正态分布. 本文所采集的吸毒人员数据变换前、后的偏度如表2所示.

表2 特征变量的偏度

Table 2 The skewness of characteristic variables

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
变换前偏度	0.10	2.62	−1.00	0.34	0.75	−0.62	−0.03	0.17	1.41	2.92
变换后偏度	0.10	−0.09	−1.00	−0.02	0.01	−0.31	−0.03	−0.04	0.19	0.00

1.3 标准化处理

对于模型来说,不同特征变量的量纲不同,模型可能会更加偏向于量纲较大的特征变量,只有经过标准化处理,才能保证模型中的每个特征变量受重视程度一样[7]. 对数值型特征变量进行标准化处理,经过Box-Cox变换前后的数据均值、标准差如表3所示.

表3 特征变量均值、标准差

Table 3 Mean and standard deviation of characteristic variables

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
均值	变化前	0.22	1.86	−0.02	−1.16	−0.61	−0.35	0.45	6.02	−1.19	−2.08
	变换后	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
标准差	变化前	0.03	0.83	0.08	0.09	0.17	0.02	0.08	0.67	0.33	0.43
	变换后	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

2　随机森林判别模型构建与分析

单一决策树存在误差较大以及过拟合的风险,为解决决策树存在的问题,Breiman 于 2001 年提出了随机森林算法(random forest,RF). 该算法采用构建一个森林,森林由多棵决策树构成,将森林中的每一棵决策树对样本的判别结果进行统计,得票数最多的类即为该样本的分类. 随机森林算法作为一种高效的分类判别方法,已广泛应用于各个领域. 本文采用脉搏波数据构建随机森林模型,进行吸毒人员的快速甄别.

2.1　模型构建

运用预处理后的 11 项数据,采用随机森林算法构建吸毒人员的甄别模型,具体步骤如下[8-9]:

(1)确定单个决策树的特征子集的特征变量的个数 m,分别设置为 3,5,7.

(2)确定森林树的棵数 n,分别设置为 10,50,100,500.

(3)计算特征子集个数和树的棵数的笛卡尔乘积,得到参数组合[m,n].

(4)用每一组参数拟合随机森林模型,共得到 3 * 4 = 12 个随机森林模型.

(5)计算得到每个随机森林模型的袋外估计的精度,选择精度最高的参数作为最优参数组合.

(6)用最优参数组合和全部数据拟合随机森林模型.

全部参数组合下,所有随机森林模型袋外估计的准确率如表 4 所示.

表 4　随机森林模型袋外估计的准确率

Table 4　The out-of-bag accuracy of Ramdom Forest models

参数组合	[3,10]	[3,50]	[3,100]	[3,500]	[5,10]	[5,50]	[5,100]	[5,500]	[7,10]	[7,50]	[7,100]	[7,500]
准确率	85.62%	86.00%	85.81%	85.56%	85.82%	85.50%	85.73%	85.69%	84.93%	85.12%	85.12%	85.19%

注:表中[a,b]参数的含义:a 代表特征变量个数,b 代表随机森林中树的棵数.

表 4 数据表明,当特征子集的特征变量个数是 3,随机森林中树的棵数是 50 时,随机森林模型的袋外估计准确率达到最大,即 86.00%.

2.2　模型测试结果分析

2.2.1　模型性能参数说明

以 TP 表示预测为正、实际为正的样例数量,FP 表示预测为正、实际为负的样例数量,TN 表示预测为负、实际为负的样例数量,FN 表示预测为负、实际为正的样例数量.

(1)准确率 = (TP+TN)/(TP+TN+FN+FP);在所有测试的人员中,预测结果与实际一致的人数的百分比.

(2)查准率 P = TP/(TP+FP);在所有预测为吸毒的人中,实际吸毒人的百分比.

(3)查全率 R = TP/(TP+FN);在所有实际吸毒的人中,成功预测为吸毒人的百分比.

(4)F1 = (2 * P * R)/(P+R). F1 值(F1 Score),是统计学中用来衡量二分类模型精确度的一种指标,它同时兼顾了分类模型的查准率和查全率[10],其最大值是 1,最小值是 0. F1 值可以看作是模型查准率和查全率的一种加权平均,是对模型性能进行评价的更全面的指标.

(5)ROC(receiver operating characteristic curve)曲线是以真阳性率(灵敏度)为纵坐标、假阳性率(特异度)为横坐标绘制的曲线. ROC 曲线越靠近左上角,试验的准确性就越高. 将各试验的 ROC 曲线绘制到同一坐标中,可直观地鉴别优劣. AUC(area under curve)是 ROC 曲线下的面积. AUC 值越大,则算法的分类性能越好.

2.2.2　模型性能参数分析

根据上述结果已知,参数组合为[3,50]的随机森林模型的准确率达到 86%,对模型进一步分析发现,其查准率为 73.61%,查全率为 40.00%,F1 为 51.84%,对应的 ROC 曲线如图 1 所示,AUC 的值为 0.87.

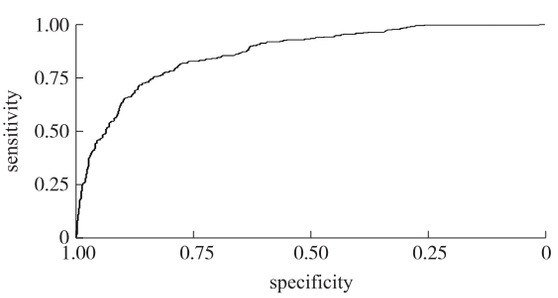

图 1　参数组合[3,50]的随机森林模型的 ROC 曲线

Fig. 1　The ROC curves of the Random Forest model with the parameter combination[3,50]

模型的相关性能参数表明,利用随机森林算法构造吸毒人员判断模型的准确率虽然较高(为 86%),但

是考虑到数据集的正负样本比例在1∶4左右,较高的准确率并不能说明模型的预测性能优异. 由于负样本的数量远多于正样本的数量,导致:(1)模型更倾向于将正样本预测为负样本,查全率只有40%的结果表明确实存在此类问题;(2)模型更加侧重于对负样本做出正确预测,而对正样本的预测效果不好,73.61%的查准率表明确实存在这种问题;(3)模型的F1值为51.84%,表明构建模型的综合性能不佳.

2.3 模型构建问题分析

随机森林算法具有较好的泛化性能且不易过拟合、对噪声不敏感等优点[11],但这样构建的随机森林模型存在如下明显问题:

(1)只依赖于袋外估计结果进行评价. 随机森林有一个重要的优点,即:在建模过程中采用袋外估计方法建立误差的无偏估计,不需要进行交叉验证或者用独立的测试集进行误差计算. 但根据文献[12],在数据集较小、难以有效划分训练集和测试集时,采用袋外估计方法虽然很有效,但是这样有可能导致乐观估计模型的泛化能力.

(2)在样本不平衡时,模型评判结果倾向于多数类样本. 随机森林模型构建过程中,对每一个样本的重视程度是一样的,损失函数的构建分配给每一个样本同样的权重. 由于正负样本的权重一样,当负样本数远超过正样本数时,模型的损失函数优化过程中,模型会更加倾向于将样本判断为负样本,因此,样本不平衡时模型的性能就会降低. 这就导致传统的分类器容易偏向于多数类,导致少数类的分类效果较差[13-15]. 在现实生活中,吸毒人员远远少于非吸毒人员,属于少数类. 上述模型没有对少数类的问题进行专门处理.

(3)模型评估以准确率为指标,过于单一. 随机森林模型构建过程中,以准确率为评价模型的指标,会造成评估指标过于单一,模型适用性不高.

综上所述,现有算法构建的模型存在较大缺陷,需进行改进以提高其实际应用价值.

3 改进随机森林模型构建及对比分析

针对上述随机森林算法构建模型存在的主要问题,从以下三方面进行改进:(1)针对只依赖于袋外估计结果进行性能评估的改进,进行训练集、测试集划分做交叉验证;(2)针对模型性能评估指标单一的改进,在最优参数组合选择的过程中采用不同的指标;(3)针对样本数据不平衡导致模型偏向于多数类判别的改进,运用下采样算法来构建平衡的训练集进行模型拟合.

3.1 模型构建的改进

3.1.1 训练集和测试集划分

为了得到较为真实的模型泛化性能的评估,对原始数据进行3次训练集和测试集的划分,设定训练集和测试集的划分比例为3∶1. 表5~表9中的袋外估计参数值是指在训练集上的袋外估计结果,与原算法的全部数据的袋外估计不同.

3.1.2 最优参数组合选择

(1)分别计算3组训练集袋外估计的F1值,以平均量为评估指标,对参数组合进行第一轮筛选. 全部参数组合下的随机森林模型,其袋外估计的F1值如表5所示.

表5 第一轮筛选模型袋外估计的F1值

Table 5 The out-of-bag F1 scores for the first round of parameter selection

参数组合	[3,10]	[3,50]	[3,100]	[3,500]	[5,10]	[5,50]	[5,100]	[5,500]	[7,10]	[7,50]	[7,100]	[7,500]
F1值	48.95%	50.12%	49.87%	50.10%	52.53%	52.19%	50.16%	52.59%	50.23%	49.95%	49.83%	49.53%

F1的最大值为52.59%,标准差为1.22%,因此低于最大值1.5个标准差范围为50.76%~52.59%,候选参数组合有[5,10],[5,50]以及[5,500]进入下一轮筛选.

(2)分别计算3组训练集袋外估计的准确率,以平均值为评估指标,对参数组合进行第二轮筛选. 其袋外估计的准确率如表6所示.

表6 第二轮筛选模型袋外估计的准确率

Table 6 The out-of-bag accuracy for the second round of parameter selection

参数组合	[5,10]	[5,50]	[5,500]
准确率	85.84%	85.57%	86.12%

准确率最大值为86.12%,标准差为0.27%,因此低于最大值1.5个标准差范围为85.71%~86.12%,因

此候选参数组合有[5,10]和[5,500],进入下一轮筛选.

(3)分别计算 3 组训练集袋外估计的 AUC 值,以平均值为评估指标,对参数组合进行第三轮筛选. 其袋外估计的 AUC 值如表 7 所示.

AUC 值最大为 0.849 1,标准差为 0.001 4,因此低于最大值 1.5 个标准差范围为 0.847 0~0.849 1,因此候选参数组合有[5,10]和[5,500],进入下一轮筛选.

(4)分别计算 3 组测试集的 F1 值,以平均值作为评估指标,对参数组合进行第四轮筛选. 在测试集上的 F1 值如表 8 所示.

表 7　第三轮筛选模型袋外估计的 AUC 值

Table 7　The out-of-bag AUC value for the third round of parameter selection

参数组合	[5,10]	[5,500]
AUC 值	0.847 1	0.849 1

表 8　第四轮筛选模型在测试集上的 F1 值

Table 8　The out-of-bag F1 score for the fourth round of parameter selection

参数组合	[5,10]	[5,500]
F1 值	0.494 6	0.526 3

表 8 表明,当特征子集的特征变量的个数为 5,森林中树的棵数为 500 时,能够拟合出性能最好的随机森林模型.

3.1.3　样本平衡化的改进

由于是在原数据集上确定最优参数,依旧没有解决正负样本不平衡的问题,因此利用确定的最优参数,在平衡的数据集上完善构建最终的随机森林模型. 由于脉搏波数据的样本分布不平衡,且正例样本的数量不算太少,本文采用下采样算法来构建平衡样本数据[16-17]. 具体操作流程如下:

(1)运用前述已划分出的训练集和测试集.

(2)在训练集中将吸毒人员和正常人员分开,构建吸毒训练样本、正常训练样本.

(3)在正常训练样本中进行随机采样,得到的样本数量和吸毒训练样本数量一致.

(4)将随机采样的正常人群样本和吸毒训练样本合并,得到新的样本类别平衡的训练集.

(5)在新的训练集进行模型拟合.

3.2　模型的拟合结果

在合成的平衡训练集中,采用已选定的参数组合进行改进的随机森林模型拟合. 3 组训练集上袋外估计的准确率、查准率、查全率以及 F1 值如表 9 所示.

从表 9 可发现,各项参数值都比较理想,没有明显弱项,意味着正例、负例的判别效果较为均衡.

表 9　3 组训练集模型拟合结果的相关参数

Table 9　Relevant parameters of fitting results of three training sets

	准确率	查准率	查全率	F1
训练集 1 袋外估计	74.17%	73.58%	75.42%	74.49%
训练集 2 袋外估计	75.00%	75.21%	74.58%	74.90%
训练集 3 袋外估计	78.54	78.66%	78.33%	78.50%

3.3　模型测试对比

利用前述划分所形成的训练集和测试集,对改进前后两个模型进行性能评价. 原模型是原始数据集拟合生成的,改进模型是用平衡训练集拟合生成的.

表 10 是两个模型在 3 组测试集上的准确率、查准率、查全率以及 F1 值. 图 2 是两个模型在测试集 1 上的 ROC 曲线及 AUC 值.

表 10　改进前后两个模型在 3 个测试集上的性能参数对比

Table 10　Comparison between performances of the two models on three test sets

		准确率	查准率	查全率	F1
测试集 1	改进前	84.59%	66.67%	36.67%	47.31%
	改进后	82.39%	52.13%	81.67%	63.64%
测试集 2	改进前	86.79%	73.68%	46.67%	57.14%
	改进后	81.45%	50.50%	85.00%	63.35%
测试集 3	改进前	83.33%	60.61%	33.33%	43.01%
	改进后	82.70%	52.81%	78.33%	63.09%

表 10 表明,原模型的准确率、查准率虽然略高于改进模型的准确率、查准率,但改进模型的查全率和 F1 值均明显高于原模型. 而实际工作中要实现对吸毒人员的甄别,则查全率和 F1 值这两个参数更为重要,所以改进的模型更符合实际需要.

图 2 的结果表明,改进模型的 AUC 值略高于原模型的 AUC 值,模型性能有所提高. 另外,两个模型在测试集 2、测试集 3 的 ROC 曲线及 AUC 值与在测试集 1 的图形类似,本文不作赘述.

综上表明:针对原模型偏向于对负类样本判别的问题,改进模型有较大改善. 在吸毒人员为少数类的这种不平衡数据集中,更加注重的是能综合体现查准率和查全率的指标 F1 值,因此,改进模型的性能要明显优于原模型的性能,能更准确地甄别出吸毒人员.

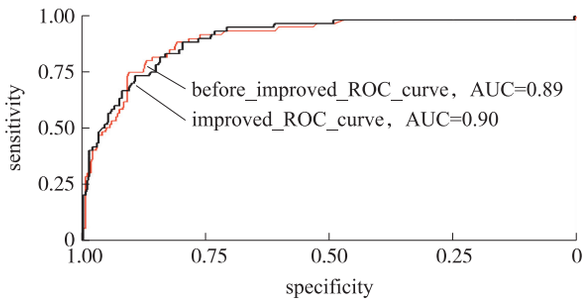

图 2　两个模型在测试集 1 的 ROC 曲线及 AUC 值

Fig. 2　The ROC and AUC of the two models on test set 1

4　结论

本文研究了利用随机森林算法构建吸毒人员甄别模型的实现过程,包括采集脉搏波数据的预处理方法,随机森林甄别模型的构建、存在的问题,改进模型的构建. 通过对改进前后两个模型进行的性能对比发现,利用脉搏波数据构建的吸毒人员随机森林甄别模型,虽然具有较高的准确度,但是整体性能不好;对随机森林算法进行改进而构建的改进随机森林模型,其查全率、F1 值等性能得到了较大提升,从而为实现基于脉搏波的吸毒人员的甄别提供了一种更可靠的建模方案.

[参考文献]

[1] 顾海艳,林祝发. 基于 Healthme 采集分析系统的吸毒者脉搏波特征研究[J]. 中国人民公安大学学报(自然科学版),2018,24(3):25-29.

[2] 李淑娟,张发祥,倪家升,等. 基于 FBG 的触点式动态压力传感器及其在脉象信息测量中的应用[J]. 光电子·激光,2016,27(10):1017-1022.

[3] MCGARRY M, NAULEAU P, APOSTOLAKIS I, et al. In vivo repeatability of the pulse wave inverse problem in human carotid arteries[J]. Journal of biomechanics, 2017, 64: 136-144.

[4] REECE A S, NORMAN A, HULSE GK. Cannabis exposure as an interactive cardiovascular risk factor and accelerant of organismal ageing: a longitudinal study[J]. Bmj Open, 2016, 6(11): e011891.

[5] 张琳. 关于居民消费水平影响因素及影响水平的实证分析[J]. 统计与管理,2017(6):69-71.

[6] 章溢,龚海林. 偏度系数的近似线性贝叶斯估计[J]. 统计与决策,2017(10):78-81.

[7] 顾爱华. 云计算网络中高维数据标准化处理优化仿真[J]. 计算机仿真,2017,34(3):317-320.

[8] 赵发林,张涛,李康. 基于遗传算法的随机森林模型在特征基因筛选中的应用[J]. 中国卫生统计,2016,33(4):559-562.

[9] 陈煜,周继恩,杜金泉. 基于交易数据的信用评估方法[J]. 计算机应用与软件,2018,35(5):168-171.

[10] 黄浩,徐海华,王羡慧,等. 自动发音错误检测中基于最大化 F1 值准则的区分性特征补偿训练算法[J]. 电子学报,2015,43(7):1294-1299.

[11] 李婉华,陈宏,郭昆,等. 基于随机森林算法的用电负荷预测研究[J]. 计算机工程与应用,2016,52(23):236-243.

[12] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[13] 孙宽宏. 不平衡数据分类方法研究[D]. 西安:西安电子科技大学,2015.

[14] 尹华,胡玉平. 基于随机森林的不平衡特征选择算法[J]. 中山大学学报(自然科学版),2014,53(5):59-65.

[15] 姚登举,杨静,詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版),2014,44(1):137-141.

[16] 胡小生,温菊屏,钟勇. 动态平衡采样的不平衡数据集成分类方法[J]. 智能系统学报,2016,11(2):257-263.

[17] 杨毅,卢诚波,徐根海. 面向不平衡数据集的一种精化 Borderline-SMOTE 方法[J]. 复旦学报(自然科学版),2017,56(5):537-544.

[责任编辑:陆炳新]