基于机器学习的失信医疗信息预防与监测识别技术研究

高晓娟
（南京医科大学附属儿童医院 信息科，江苏 南京 210008）

摘要：针对医疗体系中对患者违约行为约束力弱、优质医疗资源浪费严重的问题。文中基于机器学习算法对失信医疗信息防御和监测机制进行了研究，设计了基于随机森林（RF）算法的失信识别模型。文中设计的随机森林算法基于CART树，使用基尼系数作为节点的判别分类标准，提升了传统决策树的收敛速度。由于随机森林算法基于集成学习的Bagging思想，可以有效避免训练过程的过拟合现象，提升了单一决策树算法的分类精度。数据测试结果表明，相较于逻辑回归、K-邻近等其他机器学习算法，该模型在分类精度上分别可以提升1.3%和1.4%，可以为社会医疗信用体系的建立与完善提供技术支撑。

关键词：医疗信用体系；机器学习；随机森林；CART
中图分类号：TN98；TP311 **文献标识码：**A **文章编号：**1674-6236（2020）17-0001-05
DOI：10.14022/j.issn1674-6236.2020.17.001

Research on the technology of prevention and identification of dishonest medical information based on machine learning

GAO Xiao-juan
（*Information Department*, *Children's Hospital of Nanjing Medical University*, *Nanjing* 210008, *China*）

Abstract: In view of the weak binding force on patients´ breach of contract and the serious waste of high-quality medical resources in the medical system. In this paper, based on machine learning algorithm, the mechanism of information defense and monitoring of dishonest medical treatment is studied, and a model of dishonest recognition based on random forest (RF) algorithm is designed. The random forest algorithm designed in this paper is based on cart tree and uses Gini coefficient as the classification standard of node discrimination, which improves the convergence speed of traditional decision tree. In addition, because the stochastic forest algorithm is based on the bagging idea of integrated learning, it can effectively avoid the over fitting phenomenon in the training process and improve the classification accuracy of single decision tree algorithm. The data test results show that compared with other machine learning algorithms such as logistic regression and K-proximity, the model can improve the classification accuracy by 1.3% and 1.4% respectively, which can provide technical support for the establishment and improvement of social medical credit system.

Key words: medical; machine learning; random forest; CART

随着我国医疗事业的发展，医疗市场愈发复杂。建立健全的医疗信用体系是规范医疗市场的重要手段之一。我国医疗市场缺乏对参与者的规范手段，导致挂号违约等失信行为频出，严重浪费了有限的医疗资源。为加大医疗行业对市场参与者的管理

收稿日期：2019-11-29 **稿件编号：**201911235
基金项目：国家自然科学基金（61372071）
作者简介：高晓娟（1974—），女，陕西西安人，硕士，工程师。研究方向：信息技术在医院中的应用研究。

力度，提高市场的准入门槛与管理水平，本文对医疗中存在的失信行为进行研究。近年来，计算机技术的发展，使得数据挖掘与机器学习技术有了长足的进步。以海量的数据为支撑，借助机器学习算法，对数据进行有效利用，提升数据价值。当算法的场景为医疗领域时，可以借助医疗市场参与者相关历史行为数据，预测其是否存在失信风险，辅助医疗市场管理人员进行决策[1-4]。

基于以上分析，本文对决策树与随机森林算法进行研究，由于随机森林算法基于集成学习思想，有效避免了训练数据集中的噪声影响，因此不会出现过拟合现象。仿真结果也表明，该方法在失信行为识别时，相较于逻辑回归与K-邻近算法有更优异的性能，对医疗信用体系建立具有重要的参考意义[5]。

1 理论基础

1.1 决策树算法

近年来，在机器学习领域，决策树是应用最广泛的算法之一。相较于神经网络算法，决策树具有灵活、可解释性强的特性。灵活性体现在决策树可以根据算法设计者的意愿进行树结构的修剪，提升算法的性能；可解释性体现在决策树的每个根节点进行决策时有置信度极高的决策标准。决策树由内部节点、有向边与叶节点这3个部分组成，基本流程如图1所示[6]。

从图1可以看出，对于一个分类树，可以实现将一个数据集D标记根据不同的特征维度A，分成不同的分类C_i。分类树分类时，不同的分类树算法依据不同节点分类标准。在图1中，依据的是信息增益，信息增益的定义方法如下。

首先定义信息熵，对于数据集D，其信息熵的定义如下：

$$Info(D)=-\sum_{i=1}^{I}\frac{|C_i|}{D}\log_2\frac{|C_i|}{D} \tag{1}$$

不同的特征对于D存在条件信息熵$InfoA(|D|)$：

$$\begin{aligned}Info_A(|D|)&=-\sum_{k=1}^{K}\frac{|D_k|}{D}\times Info(D_i)\\&=-\sum_{k=1}^{K}\frac{|D_k|}{D}\sum_{i=1}^{I}\frac{|D_{ki}|}{D}\log_2\frac{|C_{ki}|}{D}\end{aligned} \tag{2}$$

此时，特征的信息增益为：

$$Gain(A)=Info(D)-Info_A(|D|) \tag{3}$$

在学术上，不同类别的决策树使用不同的节点决策标准。本文还使用了基尼系数作为决策标准，此时的决策树被称为CART。对于多分类问题，当存在K个不同的类别时，记p_k为当前样本为k类别的概率，此时依据概率分布可以定义基尼系数：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k) \tag{4}$$

存在样本集合D时，该样本的基尼系数可以写作：

$$Gini(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{5}$$

样本集合D根据特征A被分割为不同子集D_1与D_2：

$$\begin{aligned}D_1&=\{(x,y)\in D|A(x)=a\}\\D_2&=D-D_1\end{aligned} \tag{6}$$

根据特征A划分的基尼系数为：

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \tag{7}$$

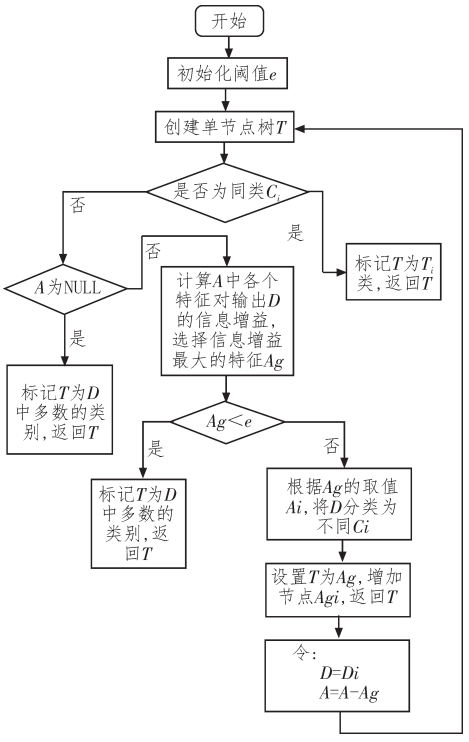

图1 决策树算法流程

1.2 集成学习与随机森林

在实际工程应用中，单个决策树的能力是有限的。训练数据中存在的噪声与孤立点会导致决策树的过拟合，严重影响决策树对未知数据分类的精度。因此，决策树生成后还需要进行剪枝操作。为了避免过拟合现象，还可以使用随机森林（RF），即组合多个单一决策树。随机森林建立要依靠集成学

习思想(Bagging)的指导[7-11]。

Bagging的特点在于样本与分类器的随机采样，其包括选取样本进行模型训练和基于分类器进行分类两个步骤，流程如图2所示。

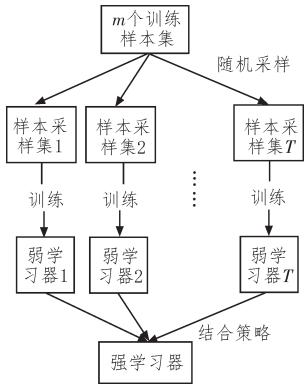

图2　Bagging算法流程

从图2看出，集成学习的难点在于对样本的随机采样与不同学习器间结合策略的设计。

本文在进行随机采样时，使用Booststrapping方式抽取训练样本T，抽取共分为k轮。k轮抽取后，可以得到k个训练集。原始数据中未被抽取的概率为：

$$\lim_{N->\infty}\left(1-\frac{1}{N}\right)^{N}\approx 0.368 \tag{8}$$

随机生成训练集后，进行决策树的生成，k个训练集可以训练出k个决策树。在训练的过程中，不需要对决策树进行剪枝操作。训练完成后，可以得到k个弱分类器。随后，对于不同的分类器赋予相同的权值，得到强分类器。

为衡量随机森林的分类性能，需要定义随机森林的间隔函数：

对于分类器$\{h_1(x),\ h_2(x),\cdots,h_N(x)\}$，存在样本集$\{X,Y\}$，其分布为$X,Y$：

$$mg(x,y)=av_k I(h_k(x)=y)-\max_{j\neq y} av_k I(h_k(x)=j) \tag{9}$$

其中，间隔函数可以定义为样本正确分类的平均与分类错误的平均。

对于随机森林算法，间隔函数可以记为$\mathrm{var}(mr)$，其上界可以由下式给出：

$$\mathrm{var}(mr)\leqslant\frac{\overline{P}(1-s^2)}{s^2} \tag{10}$$

其中，s表示单决策树的分类强度，P表示决策树间的相关性。上式表明，随机森林的间隔函数存在上限。当提升单棵分类树强度并降低每棵树间的相关度时，可以降低这一上限。

2　方法实现

2.1　数据输入

当前医疗环境复杂，为了探索合理的失信医疗行为的识别方法，本文对患者、医院、医疗企业这三方进行了调研。然后从患者的角度出发，利用随机森林算法识别失信行为，构建健康的医疗平台。首先本文从公共数据集获取了与公民信用相关的数据集。在进行随机森林的训练时，由于涉及医疗患者的隐私，只能从国外的平台lending club进行居民信贷数据的获取。然后在每个数据中添加相关医疗记录信息，以保证数据适用于本文所需的应用场景[12-14]。

在完成数据集的搜取后，根据图3所示的数据预处理流程，对数据进行预处理。

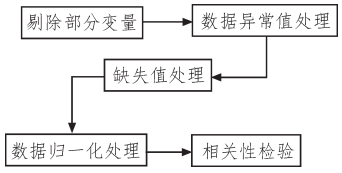

图3　数据预处理流程

在进行异常值处理时，主要是剔除严重不符合逻辑的数据项，文中采用的方法是Turkey算法。该方法可以依据数据的分布特性，定义1.5倍4分位距的数据为异常值，将其剔除。

数据归一化处理采用如下式：

$$\frac{(X-\min(X))}{\max(X)-\min(X)} \tag{11}$$

进行数据归一化后，数据本身的值将不会对评判结果造成偏移，评判结果只取决于数据属性的影响。

预处理的最后是对数据进行相关性分析，剔除数据中相关性较大的属性可以在保证分类精度的前提下，降低输入数据的维度，从而提升模型训练与分类的效率。

最终，依据图3所述的预处理步骤梳理出了图3的信用图谱，作为模型的特征输入A_1~A_{13}。

由于医疗数据涉及患者隐私，缺失较多，难以直接用于模型的训练。为了保证模型的有效性，引入了个人其他的社会征信情况，表征个人的信用状态。对于随机森林模型，每位参与医疗市场的人员都将按照图3所述的数据项建立画像。同时，依据其相关历史行为获得每个特征项的分数，最终形成模型的输入。

2.2　算法训练与测试

在随机森林中，需要根据特征向量维度、数据维

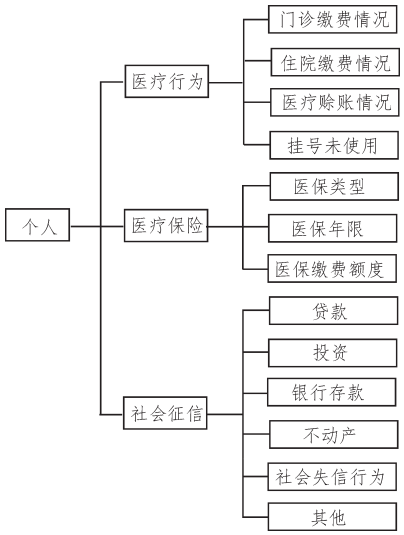

图4　个人信用图谱

度合理设置随机森林算法的相关参数[15-26]。为合理设置参数，本文对不同参数下的随机森林进行了误差分析。其的误差分析结果，如图5和图6所示。

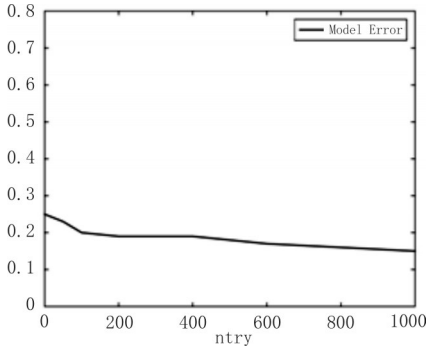

图5　ntry对模型精度的影响

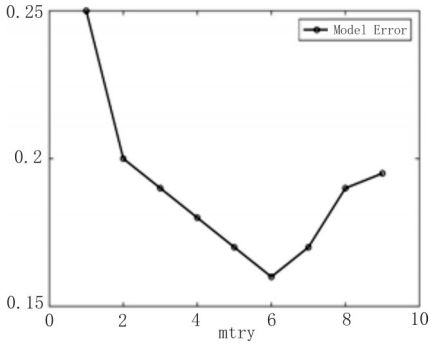

图6　mtry对模型精度的影响

图5和图6分别给出不同的ntry与mtry在训练时对于模型精度的影响。对于随机森林而言，其ntry的取值应足够大，以保证模型在训练过程中可以收敛。图5给出了在默认mtry时（mtry=sqrt(M)），不同ntry下的模型误差。由此可见，当ntry达到1 000时，模型误差降到平稳；图6给出了当ntry=1 000时，不同mtry对于模型精度的影响。由此可见，当mtry<6时，模型误差随着mtry的增大而减小；当mtry>6时，模型误差随着mtry的增大而增大。因此，mtry的最优值为6。

除了mtry与ntry，随机森林的重要参数还包括classwt。在随后的模型训练与测试中，每个参数的值如表1所示。

表1　随机森林参数设置

参数名	参数描述	参数值
ntree	树的棵树	1 000
mtry	树节点所选变量个数	6
类权重	类权重	(1,50)

在设置完模型的参数后，对模型进行了训练与测试。为了更好的衡量模型在医疗失信行为中的识别效率，本文使用逻辑回归（LR）与k-邻近算法（k-NN）进行对比。各算法的识别精度，如表2所示。

表2　不同算法的模型指标

模型	真正率	真负率	准确率
RF	12.9%	99.1%	86.6%
k-NN	1.6%	99.4%	85.2%
LR	1.6%	99.5%	85.3%

在表2中，真正率为样本为正类模型预测为正的比率；真负率为实际为负，但被模型预测为正的比率。从表2中可以看出，RF算法的真正率为12.9%，较k-NN、LR均有所提升；准确率较k-NN和LR提升分别为1.4%和1.3%，而真负率有了一定的下降。由此可见，在进行失信医疗行为识别时，RF算法有更优的性能。

3　结束语

为了更好的识别医疗行业中的失信行为，避免医疗资源浪费，本文基于机器学习与数据挖掘的思想进行相关方法的研究，基于随机森林算法建立了预防与监测模型。文中着重梳理了用于相关医疗模型的输入特征，除了结合医疗参与者的历史医疗信息之外，还引入了患者的个人社会征信状况，从而可有效弥补无医疗记录人员的行为识别与失信预防。文中使用的随机森林算法可以避免训练过程中的过拟合现象，提升预测精度。本文的内容对于医疗市场参与者的行为规范，具有一定的现实意义。

参考文献：

[1] 田臣，周丽娟.基于带多数类权重的少数类过采

样技术和随机森林的信用评估方法[J].计算机应用,2019,39(6):1707-1712.
[2] 孙悦,袁健.基于Spark的改进随机森林算法[J].电子科技,2019,32(4):60-63,67.
[3] 张鹤,张巍.基于Android的智慧医疗预约挂号客户端设计与实现[J].电子设计工程,2016,24(12):100-103,107.
[4] 刘鹏.基于Spark机器学习实现医疗保险关联频繁模式的欺诈行为挖掘技术探讨[J]. 中国数字医学,2019,14(5):15-18.
[5] 田臣,周丽娟.基于带多数类权重的少数类过采样技术和随机森林的信用评估方法[J].计算机应用,2019,39(6):1707-1712.
[6] 郑博文,赵逢禹.基于决策树分类算法异构数据的索引优化[J].电子科技,2018,31(3):48-52,60.
[7] 常晓花,熊翱.基于Adaboost的随机森林算法在医疗销售预测中的应用[J].计算机系统应用,2018,27(2):202-206.
[8] 赵锦阳,卢会国,蒋娟萍,等.一种非平衡数据分类的过采样随机森林算法[J].计算机应用与软件,2019,36(4):255-261,316.
[9] 关晓蔷,庞继芳,梁吉业.基于类别随机化的随机森林算法[J].计算机科学,2019,46(2):196-201.
[10]邓意恒,潘遂壮,欧阳少谦,等.基于"互联网+信用医疗"的先诊疗后付费模式的应用[J].中国数字医学,2019,14(5):46-48.
[11]王杰,程学新,彭金柱.一种基于粒子群算法优化的加权随机森林模型[J].郑州大学学报:理学版,2018,50(1):72-76.
[12]李秋月.基于区块链技术对医疗领域信用体系建设的研究[D].苏州:苏州大学,2018.
[13]杨兴雨,李华平,张宇波.基于聚类和随机森林的协同过滤推荐算法[J].计算机工程与应用,2018,54(16):152-157.
[14]宫芳芳,孙喜琢,罗俊霞,等.探索构建医疗信用评价体系,助力实施健康中国战略[J].现代医院管理,2018,16(6):24-26.
[15]王诚,王凯.一种基于聚类约简决策树的改进随机森林算法[J].南京邮电大学学报:自然科学版,2019,39(3):91-97.
[16]汪桂金.随机森林算法的优化改进及其并行化研究[D].南昌:南昌大学,2019.
[17] 孙悦,袁健. 基于Spark的改进随机森林算法 [J].电子科技,2019(4):60-63,67.
[18] 刘贺翔,李英娜,张长胜,等. 基于随机森林算法的Android恶意代码特征分析 [J].电子科技,2018(5):28-32.
[19] 张万福.基于随机森林的图像语义分割算法的研究 [J].电子科技,2017(2):72-75.
[20]董婷.基于随机森林算法对蛋白质相互作用的识别和预测[J]. 自动化与仪器仪表,2015(11):190-193.
[21]黄青平,邹晓明,刘楚群,等. 基于小波分解与随机森林的短期负荷预测[J]. 电力信息与通信技术,2019(9):24-29.
[22]朱龙珠; 宫立华; 刘鲲鹏,等. 基于随机森林算法的投诉预警模型优化方法[J]. 电力信息与通信技术,2018(8):24-29.
[23]杨彦荣,宋荣杰,; 胡国强,等.基于随机森林和纹理特征的苹果园遥感提取[J]. 现代电子技术,2020(3):40-44.
[24]刘云翔; 陈斌; 周子宜一种基于随机森林的改进特征筛选算法[J]. 现代电子技术,2019(12):117-121.
[25]杭琦,杨敬辉.机器学习随机森林算法的应用现状[J].电子技术与软件工程,2018(24):125-127.
[26]马晓君,董碧滢,王常欣.一种基于PSO优化加权随机森林算法的上市公司信用评级模型设计[J].数量经济技术经济研究,2019(12):165-182.