·论著·

基于机器学习的急性胰腺炎中医辨证模型构建

谈贝[1] 郑飞波[2] 张坤[3] 崔云峰[3]

[1]天津中医药大学,天津 301617;[2]天津医科大学,天津 300070;[3]天津市南开医院肝胆胰外科,天津 300100
通信作者:崔云峰,Email:nkyycyf@163.com

【摘要】 目的 基于机器学习构建急性胰腺炎(AP)的中医智能辨证模型,并比较不同机器学习算法模型的效能。**方法** 检索中国知网数据库,收集2004年12月至2022年3月公开发表的应用中医药治疗AP的文献资料,建立AP中医辨证信息数据库。运用决策树(DT)、随机森林(RF)、支持向量机(SVM)、人工神经网络(ANN)、K-近邻(KNN)5种机器学习算法构建AP中医辨证模型。采用五折交叉验证法对不同算法模型的效能进行评估;采用RF模型分析各个症状体征对于AP辨证分型的重要性。**结果** 最终纳入符合要求的中医药治疗AP的相关文献260篇。将所有特征中出现频次低于10次的症状体征或证型剔除,最终留取53个症状体征作为特征变量,获得4个AP常见证型,分别为腑实热结证、肝郁气滞证、瘀毒互结证、湿热蕴结证。分别构建不同机器学习算法的AP中医辨证模型,经五折交叉验证显示,基于RF算法的模型效果最佳,其准确率、查准率、查全率和F1分数均在95%以上(分别为96.2%、97.1%、95.6%、96.1%);而DT和KNN模型的各项效能评估结果较差。基于RF模型的特征重要性分析显示,特征重要性数值排名前10位的症状体征依次为身目发黄(0.076 8)、脉洪大(0.059 7)、苔白(0.056 7)、腹满硬痛拒按(0.053 5)、舌淡红(0.053 1)、脉弦(0.049 3)、脉涩(0.047 7)、舌质红(0.045 9)、舌有瘀斑(0.043 0)、苔薄(0.040 3)。**结论** 基于RF构建的AP中医辨证模型具有较高的准确率。

【关键词】 急性胰腺炎; 机器学习; 中医辨证模型; 随机森林

基金项目: 天津市医药卫生中医中西医结合科研项目(2021006);天津市中医药重点领域科研项目(2022005)

DOI:10.3969/j.issn.1008-9691.2023.03.014

Based on machine learning construction of traditional Chinese medicine syndrome identification model of acute pancreatitis

Tan Bei[1], Zheng Feibo[2], Zhang Kun[3], Cui Yunfeng[3]

[1]Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China; [2]Tianjin Medical University, Tianjin 300070, China; [3]Department of Hepatobiliary and Pancreatic Surgery, Tianjin Nankai Hospital, Tianjin 300100, China

Corresponding author: Cui Yunfeng, Email: nkyycyf@163.com

【Abstract】 Objective To establish intelligent traditional Chinese medicine (TCM) syndrome identification models for acute pancreatitis (AP) based on machine learning, and compare the performance of different machine learning algorithm models. **Methods** The database of China National Knowledge Infrastructure (CNKI) was researched to collect published literatures on the application of TCM for the treatment of AP from December 2004 to March 2022, and a database of TCM identification information of AP was established. Five machine learning methods such as decision tree (DT), random forest (RF), support vector machine (SVM), artificial neural network (ANN), and K-nearest neighbor (KNN) were applied to construct TCM syndrome identification models for AP. Five-fold cross-validation was used to evaluate the effectiveness of different algorithmic models. RF was applied to analyze the importance of each symptom and sign for the TCM syndrome identification of AP. **Results** A total of 260 papers related to the treatment of AP with TCM that fulfilled the requirements were finally enrolled. The symptoms and signs among all features or TCM syndrome types that occurred less than 10 times were excluded, and finally 53 symptoms and signs were retained as characteristic variables and 4 common TCM syndrome types of AP were obtained, namely, Fu-organ excess and heat retention syndrome, liver Qi stagnation syndrome, intermingled toxin and blood stasis syndrome, and dampness-heat amassment syndrome. TCM syndrome identification models for AP with different machine learning algorithms were constructed. Five-fold cross-validation showed that the model based on the RF algorithm worked best, with accuracy, precision, recall and F1 score all above 95% (96.2%, 97.1%, 95.6%, 96.1%, respectively). However, the DT and KNN models had poorer results for each effectiveness assessment. The feature importance analysis based on the RF model showed that the top 10 signs and symptoms in the ranking of feature importance were yellowing of the skin and eyes (0.076 8), flooded pulse (0.059 7), white tongue coating (0.056 7), full and stiff abdomen with pain refusing to be pressed (0.053 5), pale red tongue (0.053 1), stringent pulse (0.049 3), astringent pulse (0.047 7), red tongue (0.045 9), petechial on the tongue (0.043 0), and thin tongue coating (0.040 3). **Conclusion** The TCM syndrome identification model of AP constructed based on RF had relatively high accuracy.

【Key words】 Acute pancreatitis; Machine learning; Traditional Chinese medicine syndrome identification model; Random Forest

Fund program: Project of Scientific Research on Combination of Traditional Chinese Medicine and Western Medicine of Tianjin Medicine Health (2021006); Tianjin Science and Technology Projects in Key Areas of Traditional Chinese Medicine (2022005)

DOI: 10.3969/j.issn.1008-9691.2023.03.014

急性胰腺炎（acute pancreatitis，AP）是一种胰腺炎症性疾病，发病率和病死率均较高[1]。传统中医学虽未明确提出“胰腺”这一概念，但《医林改错》中写道，“脾中有一管，体象玲珑”，“出水道中有回血管，其余皆系水管”[2]，该描述与胰腺的解剖结构相吻合。现代医家运用中医“扶正祛邪，整体论治”的理论，综合使用内治与外治，在AP的中西医结合治疗上取得了明显疗效[3]。在人工智能兴起的浪潮中，应用机器学习等算法助力传统医学的研究与传承发展成为趋势[4]。人工智能可通过大数据分析对疾病诊断和用药提供帮助[5]。

本研究拟依据AP各种证型的症状和体征构建5种机器学习模型，包括决策树（decision tree，DT）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）、人工神经网络（artificial neural network，ANN）和K-近邻（K-nearest neighbor，KNN），以期促进AP中医辨证的智能化精准化发展。

1　资料与方法

1.1　数据来源：考虑到模型使用的广泛性，以中国知网数据库中有关中医治疗AP的文献资料作为研究对象。以“主题＝急性胰腺炎”AND“主题＝中医 OR 主题＝中药 OR 主题＝辨证 OR 主题＝方剂”为检索式，检索2004年12月至2022年3月中国知网数据库中公开发表的运用中医药治疗AP的文献数据。对症状、证型名称进行规范化处理，将同义的症状及证型合并。

1.1.1　文献纳入标准：① 患者符合AP诊断标准的文献；② 患者临床表现、中医证型分类等资料记录清晰完整的文献；③ 对于重复报告的文献，纳入较早发表的文献。

1.1.2　文献排除标准：① 综述类及动物实验类文献；② 存在明显错误的文献；③ 患者临床表现或证型描述不明确的文献。

1.2　数据处理及数据库的建立：采用Excel软件，以双人交叉判别的方法进行数据录入，建立AP中医辨证信息数据库。将文献中出现的临床表现赋值为“1”，未出现的临床表现赋值为“0”，并录入对应的证型。使用Python 3.8.2软件读取预处理过的数据，利用Pandas、Numpy数据处理模块进行数据处理。

1.3　机器学习模型建立：将数据转化为模型可以读取使用的Numpy数组形式。调用Scikit-Learn算法库中DT、RF、SVM、ANN、KNN 5种机器学习算法，构建分类器，根据输入的症状体征输出相应的证型。

1.3.1　DT：使用Scikit-Learn算法库创建DT分类器，调用网格搜索函数，搜索模型的最优参数。用于限定节点包含训练样本数的参数为“min_samples_split”，经网格搜索得到的最优参数为“5”；用于限定分支后子节点训练样本数的参数是“min_samples_leaf”，其最优参数为“1”；用于限定DT最大深度的参数为“max_depth”，其最优参数为“11”；用于计算每个节点不纯度的参数为“criterion”，其最优参数为“entropy”（即信息熵）；用于确定分支策略的参数为“splitter”，其最优参数为“best”。

1.3.2　RF：使用Scikit-Learn算法库创建RF分类器，并使用网格搜索确定模型的最优参数。参数“criterion”设置为“gini”（即基尼系数），RF中每棵DT的最大深度参数“max_depth”设置为“8”，将评估器的数量参数“n_estimators”设置为“33”；同时调用“feature_importances”计算各个症状对于预测证型的特征重要性。

1.3.3　SVM：调用Scikit-Learn算法库中的SVM，并使用网格搜索确定模型的最优参数。SVM的核函数参数“kernel”设定为“rbf”（即高斯径向基核函数），参数“C”（即浮点数）设定为“4.4”。

1.3.4　ANN：使用Scikit-Learn算法库创建多层感知机。根据筛选出的53个症状特征，设置神经网络输入层的神经元数量为53。该神经网络包含两个隐藏层，神经元数量分别为96和64；输出层的神经元数量为4，对应4种AP的常见证型。神经网络的激活函数“activation”采用“relu”函数，网络的优化器“solver”设置为“adam”，学习率“alpha”设置为“1e−5”，最大迭代次数“max_iter”设置为1 000。

1.3.5　KNN：调用Scikit-Learn算法库中的KNN分类器，使用网格搜索确定模型的最优参数。最近邻样本数“n_neighbors”的最优参数设置为“6”，权重选项“weights”设置为“distance”。

1.4　模型效能评价：因为训练集与测试集的划分对结果会产生影响，故进行五折交叉验证，取平均值，分别用准确率、查准率、查全率、F1分数等参数对模型进行评估。准确率为模型预测正确的数量占整体数量的百分比；查准率又称精确率，是在预测为阳性的样本中真正为阳性的百分比；查全率又称召回率，是真正为阳性的样本中预测为阳性的百分比；F1分数是查准率与查全率的调和平均数。

2　结　果

2.1　样本筛选：最终共纳入符合要求的中医药治

疗 AP 的相关文献 260 篇。为了保证模型的学习效果，避免信息冗余与模型过拟合，将所有特征中出现频次低于 10 次的症状体征或证型剔除。最终留取 53 个症状体征作为特征变量，获得 4 个 AP 常见证型，分别为腑实热结证、肝郁气滞证、瘀毒互结证、湿热蕴结证。

2.2 模型效能的比较（图 1）：分别对基于 DT、RF、SVM、ANN、KNN 5 种机器学习算法建立的 AP 中医辨证模型进行五折交叉验证显示，RF 模型的准确率、查准率、查全率及 F1 分数均优于其他算法模型。

2.3 特征重要性排序：RF 机器学习算法模型可以对数据集中每个特征对于预测证型的重要性进行分析，调用“feature_importances”返回特征重要性，其数值越大说明该特征对于 RF 模型越重要。结果显示（图 2），特征重要性数值排名前 10 位的症状体征依次为身目发黄（0.076 8）、脉洪大（0.059 7）、苔白（0.056 7）、腹满硬痛拒按（0.053 5）、舌淡红（0.053 1）、脉弦（0.049 3）、脉涩（0.047 7）、舌质红（0.045 9）、舌有瘀斑（0.043 0）、苔薄（0.040 3），提示上述症状体征对于 RF 模型预测 AP 证型非常重要，同时说明这些特征对于 AP 的中医辨证具有重要意义。

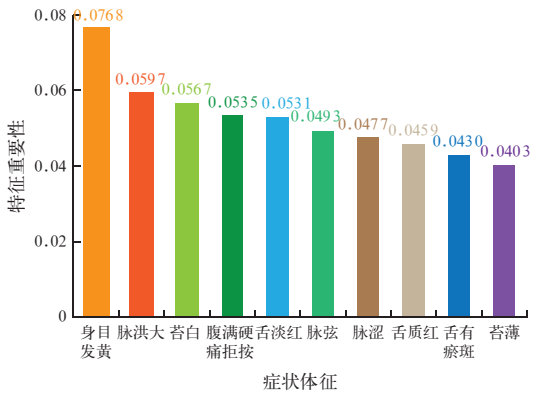
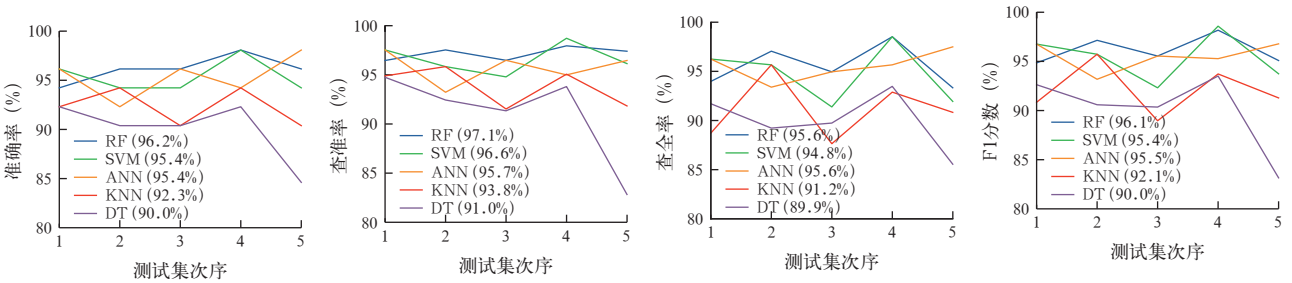


图 2 AP 中医辨证 RF 机器学习算法模型特征重要性排序

3 讨 论

本研究中以 AP 的“望闻问切”四诊信息作为“输入”，以腑实热结证、肝郁气滞证、湿热蕴结证和瘀毒互结证 4 种证型作为“输出”，采用 DT、RF、SVM、ANN、KNN 5 种机器学习算法构建 AP 中医辨证模型。结果显示，RF 在 5 种机器学习模型的效能评估中表现最佳，较其他算法模型更适用于 AP 的中医证型多分类模型构建。RF 是目前中医证型诊断模型的主流算法之一，在不孕症、类风湿关节炎等疾病的证型多分类模型或“气虚”等证素判别的二分类模型中均有应用[6-8]。分析其可能的原因在于 RF 对于高维、非线性、分散的病证数据集具有显著的处理优势；另外，每个症状或体征对证型判别的重要程度不同，RF 可在特征选择过程中对特征重要性进行计算[7]。而 DT 和 KNN 模型的各项效能较其他算法模型差，在 AP 的中医辨证数据集上表现较一般。

DT 是一种基于“if-then”规则的分层结构，从一系列有特征和标签的数据中总结出决策规则，在决策路径（边缘）中产生分叉，从而形成树状图[9]。构建 DT 是基于最优不纯度指标来找出最佳节点和分支的方法，不纯度越低，说明 DT 对训练集的拟合度就越好。衡量不纯度的计算方法包括“entropy”和“gini impurity”。RF 是以 DT 为基评估器的装袋集成算法，其预测结果是依据 DT 基评估器的预测结果进行多数表决或平均而确定的。RF 有放回地抽取训练样本，随机选择一部分特征构建模型，以减小各个 DT 之间的相关性，从而提高模型的准确性。RF 对线性及非线性数据均有良好的自适应功能，对高维数据的处理能力也较好。SVM 是通过找出边际最大的超平面作为决策边界，使分类器的误差尽可能小，因此也被称为最大边际分类器。在处理非线性数据分类的问题时，SVM 是通过非线性映



注：使用五折交叉验证法将所有数据随机分为 5 等份，依次使用其中 1 份作为测试集对模型进行评分，另外 4 份作为训练集，共进行 5 次验证及评分，取平均值

图 1 基于不同机器学习算法构建的 AP 中医辨证模型效能比较

射函数,将数据投射到高维的特征向量空间中,然后找出超平面,将数据分为两类。选用不同的核函数,可以解决不同数据分布下寻找超平面的问题。ANN也称多层感知机,是通过模拟自然神经元而建立的,它包含输入层、隐藏层和输出层。第一层是输入层,用于输入特征矩阵;最后一层为输出层,用于输出预测结果;中间的所有层均为隐藏层,在隐藏层上,每个神经元中都存在一个激活函数,下一层的神经元处理上一层神经元激活函数处理完的数据。“sklearn”中的神经网络包含3类,即随机递归神经网络玻尔兹曼机、以多层感知机为基础的神经网络分类和神经网络回归。本研究中使用的是以多层感知机为基础的神经网络分类。KNN是一种基于距离的非参数分类回归算法,其核心原理是根据训练集中特征向量空间距离待测样本最近的K个样本的多数类别作为它的分类标签。KNN算法的优点是简单易实现,但可解释性较差,对异常值不敏感,对稀有类别预测准确率也较低[10]。

基于RF得到的对于AP中医辨证分型重要性排名前10位的症状特征依次为身目发黄、脉洪大、苔白、腹满硬痛拒按、舌淡红、脉弦、脉涩、舌质红、舌有瘀斑、苔薄。脉洪大、腹满硬痛拒按是阳明腑实热结的重要表现,对于辨证具有重要意义;脉涩、舌有瘀斑提示瘀血内阻,舌质红多提示有热,综合起来有助于瘀毒互结证的诊断;脉弦是脉气紧张的表现,可以见于肝郁气滞证。肝主疏泄,调畅气机;若肝失调达,气机阻滞,则出现弦脉。身目发黄称为“黄疸”,黄色鲜明如橘属于“阳黄”,多为湿热熏蒸所致;黄色晦暗如烟熏属于“阴黄”,多为寒湿,而本次录入的文献中大多并未明确身目发黄的色泽明暗。本研究纳入的文献中出现频次较高的证型包括腑实热结证、肝郁气滞证、湿热蕴结证及瘀毒互结证,与夏庆等[11]总结的以郁、热、瘀、结病机特点为主的AP气分证相似。

近年来,越来越多的研究应用机器学习助力疾病的中医辨证论治,以及建立疾病的预测模型。曹云等[12]应用SVM、ANN和自动编码器构建了胃食管反流的中医智能辨证模型;王阶等[13]运用SVM分析了名老中医诊治冠心病的证候要素及其相应的用药规律。Lu等[5]通过收集用于治疗AP的中药复方及其疗效,应用RF构建了方剂治疗AP的疗效预测模型。机器学习在构建AP的预后模型方面也有运用,如预测重症AP患者是否发生肺损伤[14],以及AP是否会发展为重症的早期预测[15]。但是目前鲜见关于构建AP中医辨证机器学习模型的研究,本研究利用数据库中有关中医药治疗AP的文献资料,构建了机器学习智能辨证模型,为AP中医辨证的智能化精准化发展提供了思路。

本研究的局限性在于收集的数据仅来源于中国知网数据库,需要在未来广泛收集临床AP患者的四诊资料,进一步验证并完善RF中医辨证模型的效能,使其更具实用价值。

4 结　论

基于RF构建的AP中医辨证模型具有较高的准确率,将机器学习运用于中医智能辨证具有一定的应用前景。

利益冲突　所有作者均声明不存在利益冲突

参考文献

[1] Lee PJ, Papachristou GI. New insights into acute pancreatitis [J]. Nat Rev Gastroenterol Hepatol, 2019, 16 (8): 479-496. DOI: 10.1038/s41575-019-0158-2.
[2] 王清任. 医林改错 [M]. 李天德, 张学文, 整理. 北京: 人民卫生出版社, 2005.
[3] 李君秋, 肖铁刚, 曹红燕, 等. 大承气汤治疗急性胰腺炎的临床疗效观察与分析 [J]. 中华危重病急救医学, 2022, 34 (1): 91-94. DOI: 10.3760/cma.j.cn121430-20210714-01046.
[4] 项莎特, 瞿溢谦, 叶含笑. 多层学习联合建模方法设计在气阴两虚型咳嗽证候的辨证诊断中的应用 [J]. 中国卫生统计, 2020, 37 (6): 892-894. DOI: 10.3969/j.issn.1002-3674.2020.06.023.
[5] Lu WW, Chen X, Ni JL, et al. Study on the medication rule of traditional Chinese medicine in the treatment of acute pancreatitis based on machine learning technology [J]. Ann Palliat Med, 2021, 10 (10): 10616-10625. DOI: 10.21037/apm-21-2505.
[6] 许梦白, 刘雁峰, 赵宗耀, 等. 基于人工智能的不孕症中医辨证模型的构建与应用 [J]. 中华中医药杂志, 2021, 36 (9): 5532-5536.
[7] 蔡晓路. 基于随机森林的类风湿关节炎证型判别模型研究 [D]. 北京: 北京中医药大学, 2016.
[8] 舒琛洁, 梁浩, 刘淑明, 等. 机器学习算法对证候要素“气虚”辅助诊断的性能评估 [J]. 北京中医药大学学报, 2021, 44 (10): 928-934. DOI: 10.3969/j.issn.1006-2157.2021.10.010.
[9] Mitchell TM. Machine learning [M]. New York: McGraw-Hill, 1997.
[10] Zhang SC, Li XL, Zong M, et al. Efficient kNN classification with different numbers of nearest neighbors [J]. IEEE Trans Neural Netw Learn Syst, 2018, 29 (5): 1774-1785. DOI: 10.1109/TNNLS.2017.2673241.
[11] 夏庆, 黄宗文, 蒋俊明, 等. 以“益活清下”为主的中西医结合综合疗法治疗重症急性胰腺炎1 161例疗效报告 [J]. 中国中西医结合急救杂志, 2006, 13 (3): 131-134. DOI: 10.3321/j.issn:1008-9691.2006.03.001.
[12] 曹云, 卢毅, 陈建新, 等. 基于机器学习的胃食管反流病中医智能辨证模型的应用 [J]. 北京中医药大学学报, 2019, 42 (10): 869-874. DOI: 10.3969/j.issn.1006-2157.2019.10.012.
[13] 王阶, 吴荣, 周雪忠. 基于支持向量机的名老中医治疗冠心病证候要素研究 [J]. 北京中医药大学学报, 2008, 31 (8): 540-543, 560. DOI: 10.3321/j.issn:1006-2157.2008.08.010.
[14] Fei Y, Gao K, Li WQ. Artificial neural network algorithm model as powerful tool to predict acute lung injury following to severe acute pancreatitis [J]. Pancreatology, 2018, 18 (8): 892-899. DOI: 10.1016/j.pan.2018.09.007.
[15] Thapa R, Iqbal Z, Garikipati A, et al. Early prediction of severe acute pancreatitis using machine learning [J]. Pancreatology, 2022, 22 (1): 43-50. DOI: 10.1016/j.pan.2021.10.003.

（收稿日期：2022-09-27）
（责任编辑：孙茜　邸美仙）