

本文引用:孙资金, 吉 静, 马重阳, 张风君, 赵宏跃, 王雪茜, 王庆国, 程发峰. 基于机器学习的中风中医辨证模型的构建与应用[J]. 湖南中医药大学学报, 2023, 43(4): 694–699.

基于机器学习的中风中医辨证模型的构建与应用

孙资金¹, 吉 静¹, 马重阳², 张风君³, 赵宏跃⁴, 王雪茜¹, 王庆国¹, 程发峰¹

1.北京中医药大学中医学院, 北京 100029; 2.首都医科大学中医药学院, 北京 100069;

3.山东中医药大学针灸推拿学院, 山东 济南 250355; 4.哈尔滨医科大学附属第一医院, 黑龙江 哈尔滨 150007

〔摘要〕 **目的** 建立基于人工智能的中风中医辨证模型, 为中风中医智能辨证模型的构建与应用提供方法和依据。**方法** 检索中国期刊全文数据库, 收集关于中风的中医病案五种证型各 60 例, 建立中风病案中医信息数据库, 采用经过超参数调优的支持向量机(support vector machine, SVM)、K-近邻(K-nearest neighbor, KNN)、随机森林(random forest, RF)、极端随机树(extremely randomized trees, ExtraTrees)、XGBoost 及 LightGBM 对数据进行机器学习建模。全部数据的 70% 作为训练集, 30% 作为测试集, 采用五折交叉验证对模型进行评价, 以 Accuracy 作为模型优劣的评价指标, 比较模型的准确性。**结果** 中风中医四诊信息为输入变量共 55 项, 中风中医证型为输出变量共 5 项。6 种模型的拟合效果较好, Accuracy 值均在 0.85 以上; 其中 SVM 模型的准确率最高, 可达 0.95。**结论** 基于 SVM 算法模型建立的中风中医辨证模型具有较好的诊断、预测能力, 机器学习技术应用于中风中医辨证模型的构建具有方法学上的可行性。

〔关键词〕 人工智能; 机器学习; 中风; 中医辨证模型; 中医药现代化; 证候; 大数据

〔中图分类号〕 R241; TP18

〔文献标志码〕 A

〔文章编号〕 doi:10.3969/j.issn.1674-070X.2023.04.019

Construction and application of stroke TCM pattern differentiation model based on machine learning

SUN Zijin¹, JI Jing¹, MA Chongyang², ZHANG Fengjun³, ZHAO Hongyue⁴, WANG Xuexi¹, WANG Qingguo¹, CHENG Fafeng¹

1. School of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China; 2. School of Chinese Medicine, Capital Medical University, Beijing 100069, China; 3. School of Acupuncture-Moxibustion and Tuina, Shandong University of Chinese Medicine, Jinan, Shandong 250355, China; 4. The First Hospital of Harbin Medical University, Harbin, Heilongjiang 150007, China

〔Abstract〕 **Objective** To establish a TCM pattern differentiation model of stroke based on artificial intelligence, and to provide methods and basis for the construction and application of TCM intelligent pattern differentiation model of stroke. **Methods** Chinese Journal Full-text Database (CJFD) was searched for the five pattern types of TCM medical records regarding stroke, with 60 cases in each type. Then a TCM information database of stroke medical records was established. Support vector machine (SVM), K-nearest neighbor (KNN), random forest (RF), extreme random trees (Extra Trees), XGBoost, and LightGBM after hyper-parameter optimization were used to construct machine learning models, and 70% of the total data was used as the training set and 30% as

〔收稿日期〕 2022-09-13

〔基金项目〕 国家自然科学基金项目 (U21A20400); 燕京刘氏伤寒流派传承工作室项目 (1190062620029)。

〔第一作者〕 孙资金, 男, 硕士研究生, 研究方向: 人工智能与大数据、生物信息学挖掘、经典方剂的应用基础研究、经方治疗常见病疑难病。

〔通信作者〕 * 程发峰, 男, 博士, 研究员, 博士研究生导师, E-mail: fafengcheng@gmail.com。

the test set. Meanwhile, five-fold cross-validation was used to evaluate each model, and Accuracy was the evaluation index to compare the models accuracy. **Results** There were 55 input variables (information of stroke obtained by four diagnostic methods of TCM), and 5 output variables (TCM patterns of stroke). The fitting effect of the six models was good, and the accuracy values were all above 0.85; among which, the accuracy of SVM model was the highest, up to 0.95. **Conclusion** The TCM pattern differentiation model of stroke based on the SVM algorithm can diagnose and predict well, therefore, it is methodologically feasible to apply machine learning technology for constructing TCM pattern differentiation model of stroke.

[**Keywords**] artificial intelligence; machine learning; stroke; TCM pattern differentiation model; TCM modernization; pattern; big data

中风是我国常见疾病之一,其出现多伴随高血压、糖尿病等多种基础疾病,并存在着高发病率、高致残率、高死亡率、高复发率、高经济负担等特点^[1]。中医药在治疗中风方面具有悠久的历史,其病机于《黄帝内经》中即有相关记载,而方药则在东汉张仲景《伤寒杂病论》中即有论述。辨证论治是中医的特色和优势所在,其精髓之“辨”是后续临床中医选方用药的基础所在。但中医辨证多受个人主观因素影响,不确定性、不稳定性较强,常出现千人千方、辨证各不相同的情况。人工智能 (artificial intelligence, AI) 的兴起,为中医学的发展提供了新的思路。AI 技术不但可以在海量的数据中找到症状、体征与证候的相关性,并能通过知识的学习与样本数据量的扩大,发现其辨证施方的内部规律,从而深度挖掘其内部关系,促进中医现代化发展。此外,机器学习算法的应用,使得智能化、客观化辨证成为可能^[2]。基于机器学习的 AI 将中医四诊信息量化、客观化,可提高辨证的准确性和客观性^[3]。本研究基于文献资料,采用支持向量机 (support vector machine, SVM)、K-近邻 (K-nearest neighbor, KNN)、随机森林 (random forest, RF)、极端随机树 (extremely randomized trees, ExtraTrees)、XGBoost、LightGBM 分别构建中风的中医 AI 辨证模型,为中风的中医证候客观化研究提供一定依据。

1 资料与方法

1.1 资料来源

检索 2000 年 1 月至 2022 年 4 月在中国知网 (China national knowledge infrastructure, CNKI) 中公开发表的有关中风的中医病案。检索词为“中风+经验”“中风+验案”“中风+医案”“卒中+经验”“卒中+验案”“卒中+医案”,将检索词以主题方式进行检索。

1.2 诊断标准

参照《中医内科学》(中国中医药出版社,第十版,十三五规划教材)关于中风的诊断与辨证标准的相关内容,确定中风 5 个常见证型,即风痰阻络证、气虚血瘀证、阴虚风动证、痰热腑实证、风阳上扰证,以及各证型的诊断标准。

1.3 纳入标准

(1)符合中风诊断标准的文献;(2)文献中病案资料完整,包含临床表现、辨证诊断、治疗原则、治疗手段等;(3)文献发表年限为 2000 年 1 月至 2022 年 4 月;(4)对于重复发表的文献,纳入较先发表的文献;(5)治疗手段为中医治疗或中西医结合治疗;(6)文献类型为学术期刊。

1.4 排除标准

(1)文献中未标明治疗效果或治疗效果不佳;(2)报道信息太少而无法使用的文献;(3)动物实验或综述类型的文献;(4)会议论文/报刊报道/学位论文等;(5)呃逆、不宁腿综合征、卒中后抑郁等不是以中风作为主证的病例;(6)治疗手段与中医无关;(7)文献记载为纯理论论述。

1.5 分析指标

将症状、体征等中医四诊信息条目作为 AI 学习特征,证型作为分类标签。四诊信息条目共 55 项,包括神志昏蒙、精神萎靡、肢体不遂、肢体麻木、肢体困重、肢体强直、口干口渴、耳鸣耳聋、视物干涩模糊、腰膝酸软、心悸、气短乏力、纳差食少、口苦、舌红、舌瘀点、舌淡嫩、舌裂纹、舌齿痕、舌暗、舌黯淡、舌下络脉瘀滞、苔少苔剥、苔白、苔黄、苔腻、苔厚、脉沉、脉细、脉弱、脉缓、脉弦、脉涩、脉滑、脉大、脉数、汗出、大便溏、大便不行、大便干、小便黄、肢体抽搐、面赤、气粗气喘、喉中痰鸣、身热、口舌歪斜、急躁易怒、头沉、头痛、痰多、嗜睡、眩晕、不寐少寐、烦躁不安。证型共 5 项,包括风痰阻络证、气虚血瘀证、阴

虚风动证、痰热腑实证、风阳上扰证。

1.6 数据处理及数据库建立

由两名研究人员完成文献筛选,并将病例整理录入 Microsoft Excel 2019,建立中风病案中医信息数据库。对证候名称进行规范化处理,剔除出现次数较少的症状及体征(如脉结代、舌糙等),各证候要素进行语言规范化处理,根据信息的有或无将病案中出现的症状或体征分别赋值 1 或 0。赋值后的数据库作为数据源导入 Python 3.10.5,利用 Pandas、Numpy 等函数集合中的相关数据处理函数进行数据处理,包括检查与去除空值、进行定义值赋值、随机打乱病案顺序,对证型诊断进行独立编码等。为保证模型运算不出现偏倚,对每种证型病案保留 60 例进行模型构建与评估。

1.7 指标筛选

为增加模型实用性,提高计算效率,去除不相关噪声。使用 *Spearman* 相关分析与 *Lasso* 回归筛选,筛选去除类似或在判断证型过程中相对不重要的指标,相关系数阈值设定为 0.9,经分析与筛选后保留得出的特征被应用于模型训练与评定。

1.8 模型超参数设定

基于 python3.10.5,使用 scikit-learn 库进行机器学习模型构建,确保模型具有保证客观性及可重复性。本研究对所有模型调参后,对其超参数固定。针对机器学习模型(SVM、KNN、RF、ExtraTrees、XGBoost、LightGBM),其训练集与验证集比例设置为 7:3,选择 Accuracy 为主要评价指标,设定随机种

子,采用五折交叉进行交叉验证,其交叉验证测试集比例划分为 30%。

1.9 模型评价

本研究采用五折交叉验证,对模型进行进一步训练,选取 Accuracy 作为主要评价指标。

2 结果

2.1 文献整理结果

根据“1.1”“1.2”“1.3”“1.4”项下的资料获取规则,从“中风+经验”检索获得 877 篇文献,“中风+验案”检索获得 230 篇文献,“中风+医案”检索获得 388 篇文献,“卒中+经验”检索获得 284 篇文献,“卒中+验案”检索获得 49 篇文献,“卒中+医案”检索获得 62 篇文献。通过剔除重复文献,阅读文献题目、摘要,并进一步通读全文后筛选符合纳入标准的文献,共 333 篇。最终每类证型病案保留 60 例用于计算。

2.2 指标筛选结果

经过 *Spearman* 相关分析与 *Lasso* 回归筛选后,得到以下指标:神志昏蒙,精神萎靡,肢体麻木,肢体困重,肢体强直,耳鸣耳聋,心悸,气短乏力,纳差食少,舌红,舌瘀点,舌裂纹,舌齿痕,舌黯淡,舌下络脉瘀滞,苔白,苔黄,脉沉,脉细,脉弱,脉缓,脉弦,脉滑,脉大,大便溏,大便不行,小便黄,肢体抽搐,面赤,气粗气喘,急躁易怒,痰多,嗜睡,眩晕,不寐少寐,烦躁不安。这些指标将用于接下来的模型训练,具体相关分析与 *Lasso* 回归结果见图 1—3。

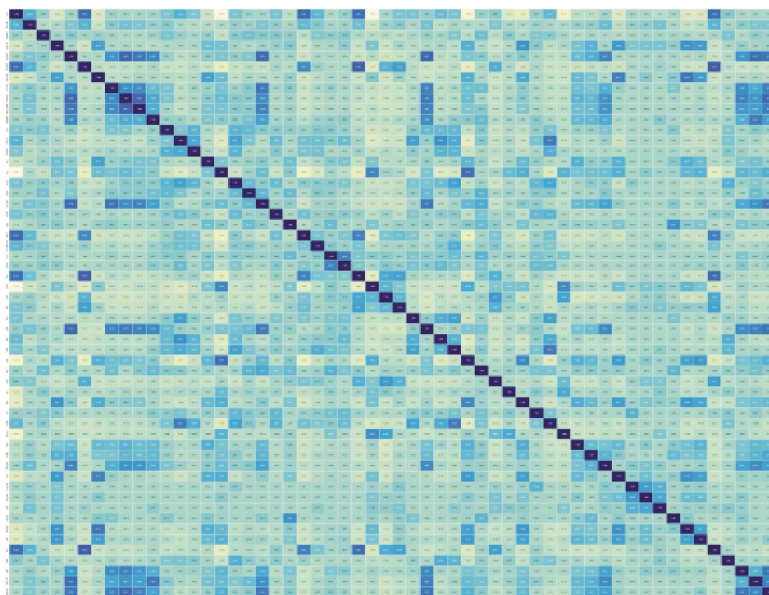


图 1 *Spearman* 相关分析热图

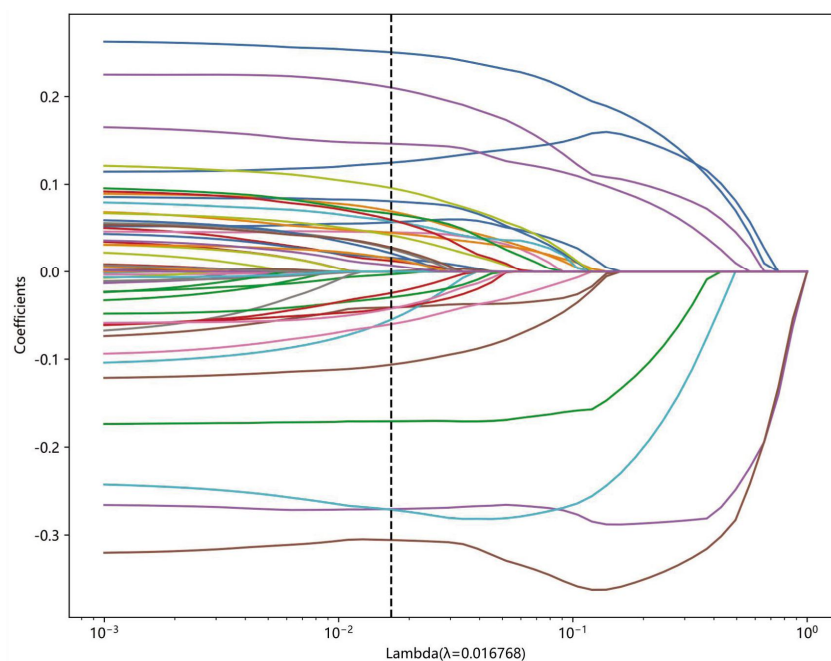


图 2 Lasso 回归系数分布图

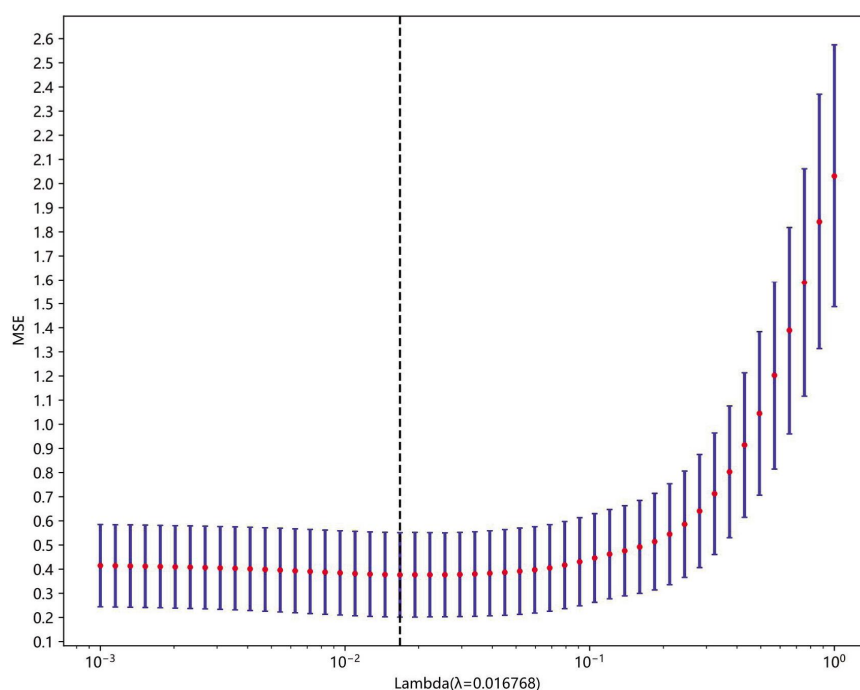


图 3 Lasso 回归阈值选择图

2.3 模型超参数设定结果

SVM 选取其中的支持向量机分类模型(support vector classification, SVC), 其中 probability=True; KNN 中 algorithm='kd_tree'; RF 中, n_estimators=10; ExtraTrees 中, n_estimators=10; XGBoost 中 base_score=None, booster=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=None, enable_categorical=False, eval_metric='error', gamma=None, importance_type=None, interaction_constraints=None,

learning_rate=None, max_delta_step=None, max_depth=None, min_child_weight=None, missing=nan, monotone_constraints=None, n_estimators=10, n_jobs=None, num_parallel_tree=None, predictor=None, random_state=None, reg_alpha=None, reg_lambda=None, scale_pos_weight=None, subsample=None, tree_method=None, use_label_encoder=False, validate_parameters=None, verbosity=None; LightGBM 中, n_estimators=10, objective='binary'。

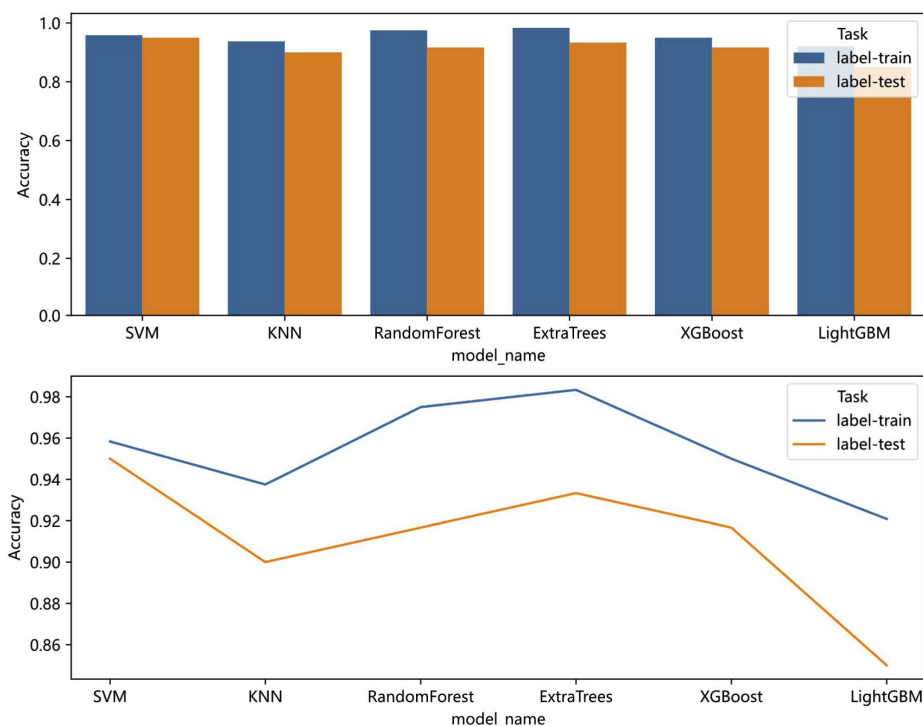


图 4 各模型 Accuracy 评价指标

2.4 模型评价结果

在验证集中得到模型验证结果如图 4 所示,其中,SVM 准确率为 0.95,KNN 准确率为 0.9,RF 准确率为 0.91,ExtraTrees 准确率为 0.93,XGBoost 准确率为 0.91,LightGBM 准确率为 0.85。SVM 模型相较其他模型而言,其准确率明显更高。

3 讨论

目前,AI 已广泛应用于医疗领域的多个方面,促进了医学大数据时代的进一步变革^[4]。在医学领域中,AI 已为影像诊断、患者预后预测与指导临床用药提供了强有力的帮助^[5-6]。

将传统的中医辨证与 AI 相结合构建智能辨证模型,尽可能减少人为的主观干预,可辅助临床诊断,是实现中医辨证诊断客观化、规范化与现代化发展的要求^[7]。如应用视觉注意机制的计算机视觉技术可为中医舌诊提供标准化和可重复化;基于中西医结合的机器学习技术可为卒中后抑郁的发病提供预测等^[8-10]。中医辨证智能化的发展需要合理的算法模型,RF、SVM、KNN、ExtraTrees、XGBoost、LightGBM 模型是目前主流应用的算法模型。

其中,支持向量机是建立在统计学习理论基础上,借助最优化方法来解决机器学习问题的新工具。它将机器学习问题转化为求解最优化问题,并应用

最优化理论来构造算法。其通过寻求最小结构化风险来提高机器学习的能力,多适用于小样本模型的训练适用^[11]。而 SVC 则是支持向量机中用于分类模型的优化算法,包含 C-SVC 与 V-SVC 两种不同的参数模型设置^[12]。KNN 算法与支持向量机算法在一定程度上存在相似性,其亦是利用训练数据对于特征向量之间表达出的不同关系来进行划分,并将划分结果作为分类模型。但因其空间复杂度高、特征维度大,故可解释性较差,且存在对于样本数量过少的类别预测准确率低等缺点^[13]。RF 是建立于决策树算法基础之上发展而来的算法,可用于机器学习的分类计算,常适用于监督学习,其利用多棵决策树对模型进行综合训练^[14]。其具有对离群值不敏感,不易对数据产生过度拟合的优点,但却具有其算法倾向于观测值较多的类别的缺点^[15]。ExtraTrees 算法与 RF 算法十分相似,都是由许多“树”构成,但该算法与 RF 的主要区别如下:(1)ExtraTrees 是使用所有的训练样本得到每棵决策树的,也就是说,每棵决策树应用的是相同的全部训练样本;(2)RF 是在一个随机子集内得到的最佳分叉属性,而 ET 是完全随机地得到分叉值,从而实现了对决策树的分叉。XGBoost 与 LightGBM 均属于 boosting 算法系列。其中,XGBoost 显式地将树模型的复杂度作为正则项加在优化目标,并允许使用列抽样来防止过拟合,借鉴了 Random

Forest 的思想,同时对树的叶子数和叶子分数做惩罚,以确保了树的简单性,实现了算力的节省^[6]。与 XGBoost 相同,作为梯度提升决策树(gradient boosting decision tree, GBDT)的一员,LightGBM 同样具有可解释性高、准确、可进行多类分类运算等优点,但其相较于 XGBoost 而言,其所具有的基于梯度的单边采样与互斥特征捆绑,使其具有更高的运算速度与更小的内存消耗,从而能够在相同的时间对更大的样本量与特征进行运算^[7]。

本文中所使用的模型准确率较高,但因文献资料与临床实际可能仍存在一定程度的偏差,且可能存在过拟合的情况,故后续仍需结合临床实际病例,对该模型进行调优。

综上所述,基于 AI 建立中风辨证模型具有一定的可行性。但基于 AI 的中风中医辨证模型尚需跨学科合作,以期在算法与临床方面得到多方位的兼顾,并在未来样本量提升与临床病例的验证后,可对其进行进一步的完善与优化,有利于提高中医辨证诊断的科学性,促进中医研究的现代化发展。

参考文献

- [1] 肖爽,朱以诚.脑卒中的性别差异:流行病学、危险因素、治疗及预后[J].中国神经免疫学和神经病学杂志,2020,27(1):57-60.
- [2] 李本岳,李伟荣,潘华峰,等.人工智能对中医诊断的影响[J].世界科学技术-中医药现代化,2020,22(5):1624-1628.
- [3] 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.
- [4] 陈梅,吕晓娟,张麟,等.人工智能助力医疗的机遇与挑战[J].中国数字医学,2018,13(1):16-18.
- [5] FARWELL M D, MANKOFF D A. Analysis of routine computed tomographic scans with radiomics and machine learning: One step closer to clinical practice[J]. JAMA Oncology, 2022, 8(3): 393-394.
- [6] 王敬瀚.ROC 曲线在临床医学诊断实验中的应用[J].中华高血压杂志,2008,16(2):175-177.
- [7] 黄欣荣,钟平玉,马纲.人工智能与中医智能化[J].中医杂志,2017,58(24):2076-2079,2106.
- [8] 许家伦,周昌乐,方肇勤,等.舌像颜色特征的计算机分析与识别研究[J].上海中医药大学学报,2004,18(3):43-47.
- [9] 刘哲,陈家旭,赵宇明,等.基于视觉注意和支持向量机的舌体自动分割方法的探讨[J].北京中医药大学学报,2013,36(1):18-20.
- [10] 罗晓舟,温小鹏,何家扬,等.基于机器学习的卒中后抑郁影响因素分析[J].中医杂志,2017,58(17):1478-1481.
- [11] CRISTIANINTNELLO. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Cambridge: Cambridge University Press, 2000: 93-124.
- [12] 汤华丽. SVM 中两类常用分类方法的关系研究[D].重庆:重庆大学,2005.
- [13] ZHANG S C, LI X L, ZONG M, et al. Efficient kNN classification with different numbers of nearest neighbors [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.
- [14] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [15] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(4):1190-1197.
- [16] 连克强.基于 Boosting 的集成树算法研究与分析[D].北京:中国地质大学,2018.
- [17] KE G L, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 3149-3157.

(本文编辑 匡静之)