

基于XGBoost和SHAP的中药寒热药性识别及寒热特征标记可视化研究*

张喜科¹, 赵文华¹, 付先军², 马志庆^{1**}, 魏国辉^{1,2**}

(1. 山东中医药大学智能与信息工程学院 济南 250355;

2. 山东中医药大学中医药经典理论教育部重点实验室 济南 250355)

摘要:目的 运用集成学习中的XGBoost算法探索构建中药寒热药性识别模型,采用SHAP进一步探索中药寒热特征标记与寒热药性识别结果的关系。方法 利用61味中药的紫外光谱数据集,运用集成学习中的XGBoost算法,构建中药寒热药性识别模型,通过稳定性评价和外推评价等方式评估模型的性能,采用SHAP对中药寒热特征标记进行可视化。结果 模型评价方面,在稳定性比较中,本文模型在单溶剂下的ACC和AUC分别为0.852、0.865,多溶剂下的ACC为0.770;在外推比较中,本文模型在单溶剂下的ACC和AUC分别为0.810、0.828,多溶剂下的ACC为0.786;可视化方面,寒性中药在紫外波长400和267 nm的吸收度具有相似性,而热性中药在紫外波长400、299和301 nm的吸收度具有相似性。结论 与经典模型相比,本文模型对中药寒热药性具有更好的识别能力。具有相似的紫外光谱的中药,其药性同样具有相似性,初步证明了物质成分相似的中药具有相似的药性。

关键词: 中药寒热药性 中药寒热特征标记 集成学习 XGBoost SHAP 可视化

doi: 10.11842/wst.20211217005 中图分类号: R285.1 文献标识码: A

中药药性理论是传统中医药的重要组成部分,中药寒、热、温、凉四性理论又是中药药性理论的核心内容之一。四性反映了中药对人体寒热变化的作用倾向,其中寒与凉、温与热只是程度上的差异并没有本质上的区别。从本质属性讲,四性可以归结为寒热两性^[1-2]。临床治疗中,寒热辨证是中医辨别疾病病因病性最基本的纲领^[3]。因此,正确识别中药寒热药性对中医辨证施治、组方遣药具有重要意义。

相关研究表明:决定中药寒热药性的主要依据是药效,而药效的基础是其所含物质成分^[4-12]。因此,中药的物质成分是产生寒热药性的基础。从中药的物质成分角度研究中药寒热药性识别主要包含两方面的内容:一是如何表征中药的物质成分;二是如何判别中药的寒热药性。近年来,中药寒热药性成为研究

热点^[13]。研究者采用中药物象^[14]、中药无机物质含量^[15]、代谢组学^[16]、三维荧光光谱^[17]、中药化学指纹图谱^[18-19]和中药化合物分子描述符^[20-22]等方式表征中药的物质成分,之后通过构建人工智能模型识别中药的寒热药性。薛付忠团队选用中药物象属性表征中药成分,使用人工神经网络(Artificial Neural Network, ANN)、主成分-线性判别分析和偏最小二乘模型识别中药寒热药性^[23-25]。王振国团队通过构建支持向量机(Support Vector Machine, SVM)模型研究发现中药寒热药性与无机物质的种类和含量具有相关性^[15]。聂斌等^[16]采用中药代谢组学数据表征中药特征,构建随机森林模型识别寒热中药。Long等^[20]、Fu等^[21]和Wei等^[22]应用分子描述符表征中药的化合物成分并构建SVM、K近邻(k-Nearest Neighbor, KNN)和极限学习机模型识

收稿日期:2021-12-17

修回日期:2022-12-13

* 国家自然科学基金委员会面上项目(81473369):基于“性-构”关系的中药成分寒热药性评价,负责人:付先军。

** 通讯作者:马志庆,教授,硕士研究生导师,主要研究方向:生物医学图像处理与分析;魏国辉,副教授,硕士研究生导师,主要研究方向:中药药性。

别中药的寒热药性。

笔者所在课题组研究发现中药化学指纹图谱可以用于识别中药寒热药性,使用中药紫外光谱数据结合距离度量算法对中药寒热药性识别进行了尝试^[18-19],但识别效果一般。在此基础上,本研究采用4种溶剂(无水乙醇、蒸馏水、氯仿、石油醚)提取中药成分的紫外光谱数据。单溶剂不能提取中药的全部物质成分,通过多溶剂提取中药的物质成分更有利于对中药寒热药性进行全面分析。但中药紫外光谱数据大多具有样本量偏小、特征偏多等特点,若继续使用已有研究采用的经典分类器,则识别效果不好。经过大量实验发现,集成学习算法可以很好地适应中药紫外光谱数据的特点,已有研究将集成学习算法应用于基于气味指纹图谱的中药鉴定,取得了不错的效果^[26]。本研究将新的集成学习算法引入基于中药紫外光谱数据的寒热药性识别,通过设计大量验证实验,从多个方面证实了集成学习算法应用于中药寒热药性识别的可行性。同时,本研究将一种全新的可视化方式引入中药寒热药性研究,通过中药寒热特征标记可视化的方式探索了中药寒热特征标记与寒热药性识别结果的关系。

1 资料与方法

1.1 极端梯度提升(eXtreme Gradient Boosting, XGBoost)算法

与经典分类算法的目标函数只使用损失函数不同,XGBoost^[27]算法的目标函数由损失函数和模型复杂度两部分组成。中药紫外光谱数据大多具有样本量偏小等特点,大部分经典分类模型应用于中药紫外光谱数据时会处于过拟合状态。模型复杂程度越高,对于训练集的识别效果越好(偏差小),但对于测试集的识别效果较差(方差大),模型的泛化能力较弱。XGBoost模型的目标函数如下所示,其中,第一部分用于衡量偏差的大小,偏差越大模型的识别效果越差;第二部分用于衡量方差的大小,模型的复杂程度越高,对于训练集的拟合越好但在测试集上的表现差异大。因此,为了达到模型的整体最优,需要综合考虑偏差和方差的影响。XGBoost算法在目标函数上的改变使其生成的模型不易发生过拟合,具有更强的泛化能力,可以更好地适应中药紫外光谱数据的特点。

$$\zeta(\varphi) = \sum_{i=1}^m l(\hat{y}_i, y_i) + \sum_{k=1}^n \Omega(f_k) \quad (1)$$

1.2 SHAP(SHapley Additive exPlanations)算法

SHAP^[28-31]是一种模型事后解释的算法,可以对复杂的机器学习模型的识别结果进行解释。本研究以61味中药的紫外光谱数据集作为模型的输入,模型的输出为中药的寒热药性。基于XGBoost算法构建的中药寒热药性识别模型在识别准确率方面表现良好,但准确率的提升带来了可解释性的下降。由于基于XGBoost算法构建的中药寒热药性识别模型几乎是一个黑箱模型,导致我们对模型的具体构建过程并不了解,使用传统的feature importance 仅能反映各寒热特征标记的重要程度,但无法了解各寒热特征标记与寒热药性识别结果的具体联系。SHAP 值的优势在于,其不仅可以反映各寒热特征标记对寒热药性识别结果的影响程度,而且可以准确指明影响程度的正负性。SHAP 值通过量化每个特征标记对模型识别结果所做的贡献来对模型的识别结果进行解释。若 x_i 表示第 i 个样本, x_{ij} 表示第 i 个样本的第 j 个特征, y_i 表示模型对该样本的识别值, y_{base} 表示模型的基线(通常为所有样本目标变量的均值),则SHAP 值服从以下等式:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) \quad (2)$$

其中, $f(x_{ij})$ 为 x_{ij} 的SHAP 值, $f(x_{ik})$ 为第 i 个样本的第 k 个特征标记对最终识别结果 y_i 的贡献值, $f(x_{ik}) > 0$ 时,表明该特征标记提升了识别值,对识别结果起正向作用;反之,则降低了识别值,对识别结果起反向作用。

1.3 数据来源

本研究选用药性明确且具有代表性的61味中药

表1 研究选用的61味中药

药性	中药				
寒性	薄荷	侧柏	柴胡	车前子	川贝母
	大黄	大青叶	淡竹叶	地肤子	防己
	甘遂	葛根	瓜蒌	海藻	黄柏
	黄连	金银花	龙胆	芦荟	络石藤
	毛知母	墨旱莲	蒲公英	秦皮	瞿麦
	生地黄	天冬	稀莩草	栀子	紫草
	白胡椒	白芥子	半夏	葶苈	补骨脂
热性	苍术	草豆蔻	陈皮	川芎	杜仲
	附子	甘松	干姜	高良姜	藁本
	红花	厚朴	麻黄	木瓜	木香
	羌活	肉桂	檀香	天南星	威灵仙
	吴茱萸	细辛	仙茅	延胡索	淫羊藿
	炙延胡索	-	-	-	-

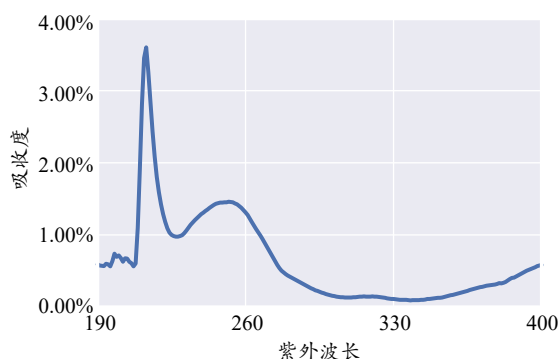


图1 薄荷(石油醚溶剂)不同紫外波长吸收度曲线

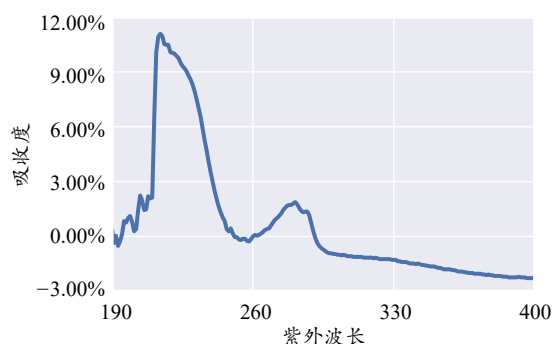


图2 干姜(石油醚溶剂)不同紫外波长吸收度曲线

表2 不同溶剂下模型识别的ACC和AUC

溶剂	ACC	AUC
无水乙醇	0.508	0.490
蒸馏水	0.590	0.689
氯仿	0.639	0.703
石油醚	0.852	0.865

(30味寒性药,31味热性药,如表1所示)。研究分别测量了在不同溶剂(无水乙醇、蒸馏水、氯仿、石油醚)中61味中药对紫外波长400–190 nm的吸收度。图1、2分别为寒性中药薄荷和热性中药干姜的紫外吸收度曲线,两曲线对比结果显示,寒热中药的紫外吸收度曲线在吸收峰的形状、所在位置及强度上存在一定差异。

1.4 模型设计

本研究的XGBoost模型基于Python语言构建,寒热特征标记可视化基于SHAP实现。

1.5 模型评价

采用准确率(Accuracy, ACC)和受试者工作特征曲线下方的面积大小(Area Under Curve, AUC)两种评价指标衡量单溶剂下模型的性能;采用ACC评价指标衡量多溶剂下模型的性能。ACC和AUC值越大,模型

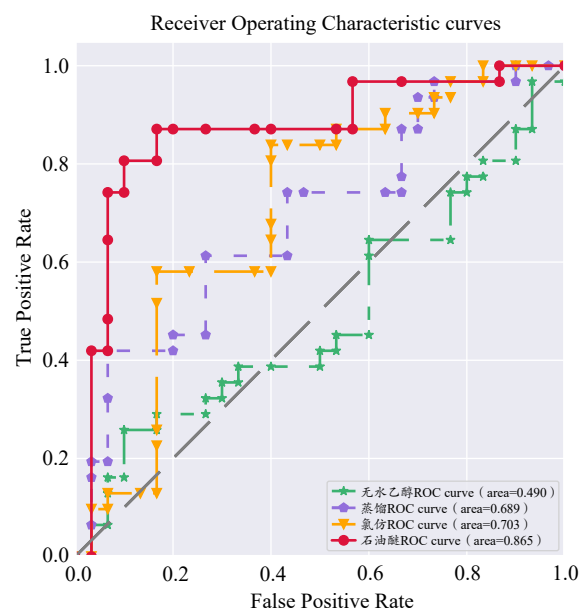


图3 不同溶剂下模型识别的ROC曲线

性能越好。

为了证明本文模型在寒热药性识别任务上的优势,设计对比实验比较本文模型与已有研究采用的ANN^[23]、SVM^[15,20]和KNN^[21]模型在上述两种评价指标上的差异。

1.5.1 稳定性评价

本研究采用留一法对模型的稳定性进行评估。留一法使用的训练集与原始数据集相比只减少了一个样本,这使得留一法中被实际评估的模型与使用原始数据集训练出的模型基本相似,评估结果比较准确。就本研究使用的61味中药紫外光谱数据而言,不同溶剂(无水乙醇、蒸馏水、氯仿、石油醚、多溶剂)下实验分别重复进行61次,每次留出1味中药样本作为测试集,剩余60味中药样本作为训练集。

1.5.2 外推评价

在模型的外推性能评估中,本研究随机将原始数据集中各约20味寒热中药样本作为训练集,剩余中药样本作为测试集。为了消除分组的偶然性对实验结果可能造成的偏差,实验重复进行10次计算平均值和方差。

2 结果

2.1 稳定性比较

不同溶剂(无水乙醇、蒸馏水、氯仿、石油醚)下模型识别的ACC和AUC如表2所示,ROC曲线如图3所

表3 单溶剂多模型稳定性比较结果

模型	ACC	AUC
人工神经网络(ANN)	0.607	0.682
支持向量机(SVM)	0.705	0.622
K近邻(KNN)	0.721	0.719
本文模型	0.852	0.865

表4 多溶剂多模型稳定性比较结果

模型	ACC
人工神经网络(ANN)	0.590
支持向量机(SVM)	0.590
K近邻(KNN)	0.754
本文模型	0.770

表5 单溶剂多模型外推比较结果

模型	ACC	AUC
人工神经网络(ANN)	0.633±0.009	0.664±0.009
支持向量机(SVM)	0.733±0.006	0.745±0.009
K近邻(KNN)	0.733±0.003	0.753±0.002
本文模型	0.810±0.004	0.828±0.004

表6 多溶剂多模型外推比较结果

模型	ACC
人工神经网络(ANN)	0.600±0.003
支持向量机(SVM)	0.667±0.005
K近邻(KNN)	0.695±0.005
本文模型	0.786±0.004

表7 寒热药性总体识别的精确率、召回率和F1值

指标	寒性中药	热性中药
精确率	0.862	0.844
召回率	0.833	0.871
F1值	0.847	0.857

表8 寒热药性总体识别的混淆矩阵

预测值 \ 真实值	寒性中药	热性中药
寒性中药	25	5
热性中药	4	27

示。实验结果表明,石油醚溶剂下的紫外光谱数据最适合用于寒热药性识别,氯仿、蒸馏水溶剂下的紫外光谱数据在寒热药性识别上的表现次之,无水乙醇溶剂下的紫外光谱数据在寒热药性识别上的表现较差,在后续多溶剂实验中,可以考虑只融合石油醚、氯仿和蒸馏水3种溶剂下的紫外光谱数据。

鉴于石油醚溶剂下的紫外光谱数据在寒热药性

识别上的表现最好,无水乙醇溶剂下的紫外光谱数据在寒热药性识别上的表现较差。因此,在设计对比实验比较4种模型的性能时,单溶剂选取石油醚溶剂下的紫外光谱数据作为原始数据,多溶剂融合石油醚、氯仿和蒸馏水3种溶剂下的紫外光谱数据作为原始数据。本文模型在单溶剂和多溶剂下与ANN、SVM和KNN模型的稳定性表现分别如表3、4所示,实验结果显示,本文模型的稳定性优于参与比较的3种模型。

2.2 外推比较

单溶剂和多溶剂下本文模型与ANN、SVM和KNN模型的外推性能表现分别如表5、6所示,实验结果显示,本文模型在外推表现上优于参与比较的3种模型。

综合表3-6的实验结果可以得出结论:与已有研究采用的经典机器学习模型相比,本文模型在单溶剂和多溶剂下均具有更好的稳定性和外推性能,更加适应中药寒热药性识别任务的需要。

2.3 寒热药性总体识别

本研究使用留一法评估模型的总体识别性能,寒热药性总体识别ACC为85.2%(52/61),寒热药性总体识别的精确率、召回率和F1值(F1值可以看作是精确率和召回率的调和平均数,是一种同时兼顾精确率和召回率的度量方式)如表7所示。表8为寒热药性总体识别的混淆矩阵,9味中药药性被误判,侧柏、柴胡、金银花、毛知母、栀子5味寒性中药被识别为热性中药,白芥子、半夏、附子、天南星4味热性中药被识别为寒性中药。通过绘制上述9味药性被误判的中药的紫外吸收度曲线(见图4),并与两种典型的寒热中药薄荷、干姜的紫外吸收度曲线进行对比,结果显示,药性被误判的9味中药的紫外吸收度曲线与其相异药性的中药的紫外吸收度曲线在吸收峰的形状、所在位置或强度上存在相似性,这是导致其药性被误判的主要原因。

2.4 寒热特征标记可视化

图5为61味中药寒热特征标记重要性排行。实验结果显示,61味中药在石油醚溶剂下对紫外波长400、301、299、256、267、373、265、226、368和251 nm的吸收度对于其寒热药性识别发挥重要作用。从整体角度来看,紫外波长400 nm的吸收度对于61味中药寒热药性识别最重要。

本研究以寒性中药薄荷、车前子,热性中药白胡椒、葶苈4种中药寒热特征标记可视化的方式进一步

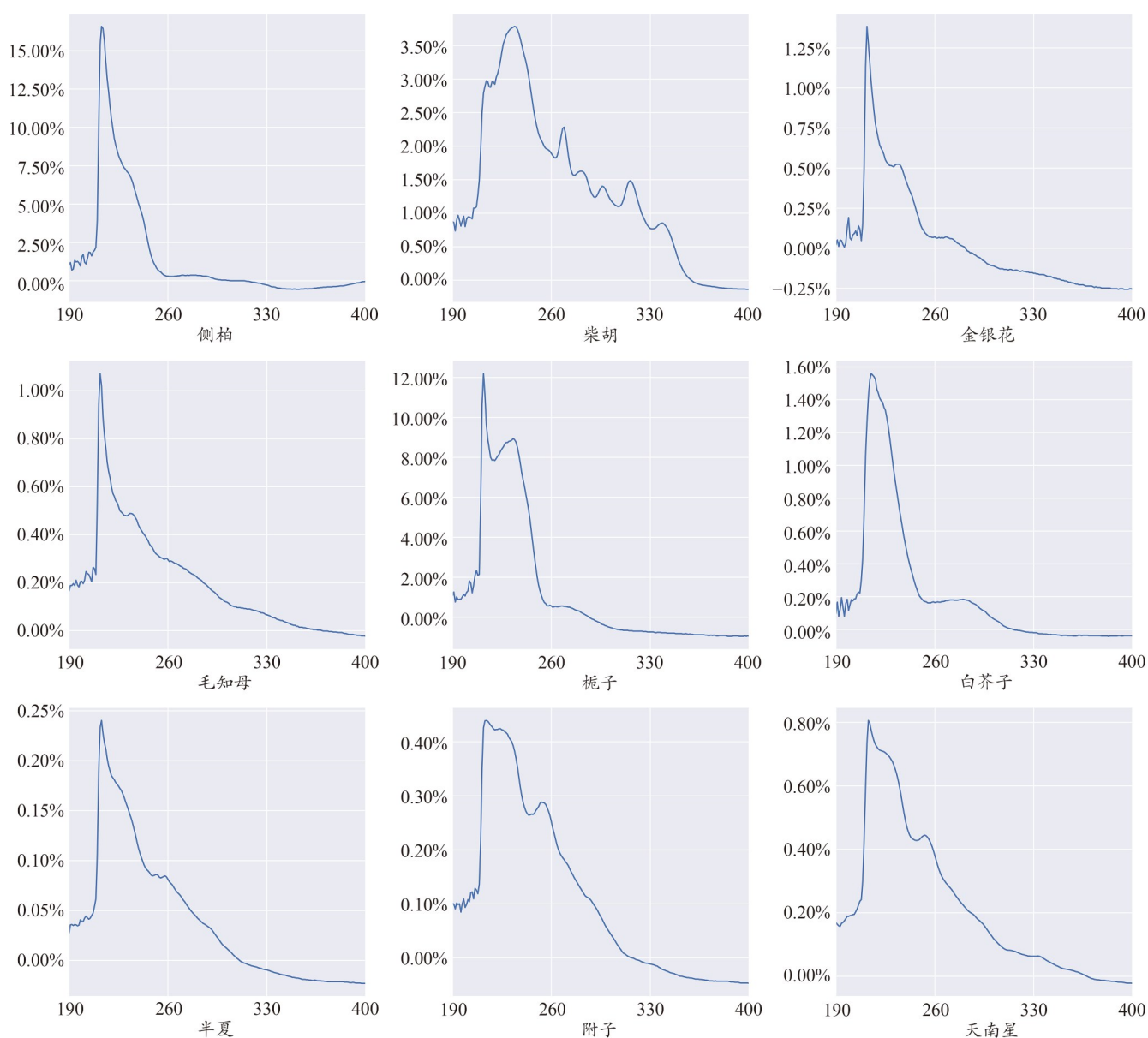


图4 药性被误判的中药(石油醚溶剂)不同紫外波长吸收度曲线

分析中药寒热特征标记与寒热药性识别结果的关系。寒热特征标记贡献度为负,则该特征标记对于该味中药为寒性特征标记;贡献度为正,则该特征标记为热性特征标记。贡献度的绝对值大小表示该特征标记对于寒热药性识别结果的影响程度大小,本研究将贡献度的绝对值大于0.5的特征标记视为重要特征标记。图6-7分别为寒性中药薄荷和车前子的寒热特征标记贡献度可视化效果图。图8-9分别为热性中药白胡椒和荜茇的寒热特征标记贡献度可视化效果图。实验结果显示,寒性中药在紫外波长400和267 nm的吸收度具有相似性,而热性中药在紫外波长400、299

和301 nm的吸收度具有相似性。

本研究通过单味中药寒热特征标记可视化的方式,分析了寒热中药在不同紫外波长吸收度上的差异。可视化的结果表明,有相似的紫外光谱的中药具有相似的药性。基于笔者所在课题组已有研究得出的结论:中药的化学指纹图谱可以有效表征中药的物质成分。本研究从中药寒热特征标记可视化的角度初步证明了:物质成分相似的中药具有相似的药性。

3 讨论

理论上讲,决定中药寒热药性的主要依据是药

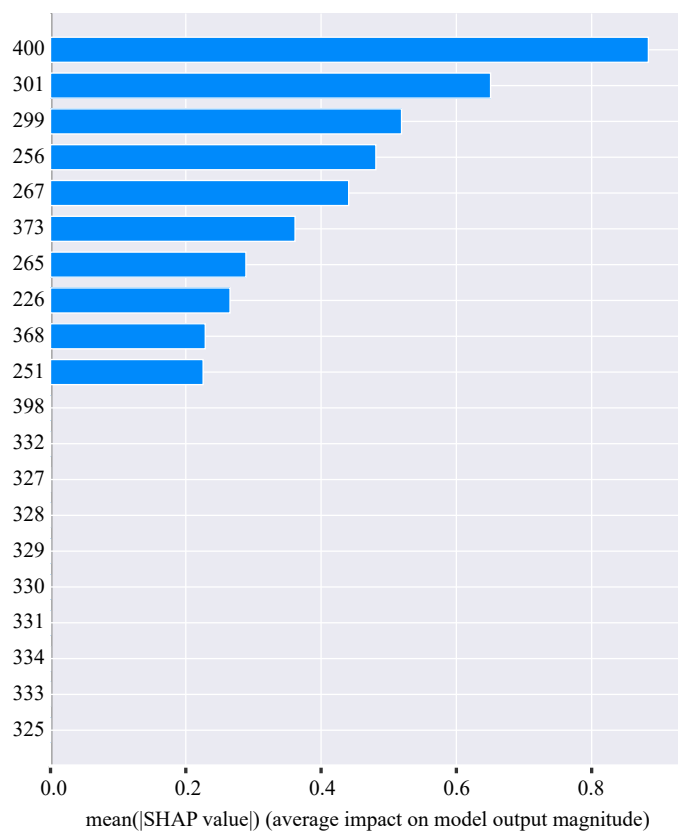


图5 寒热特征标记重要性排行

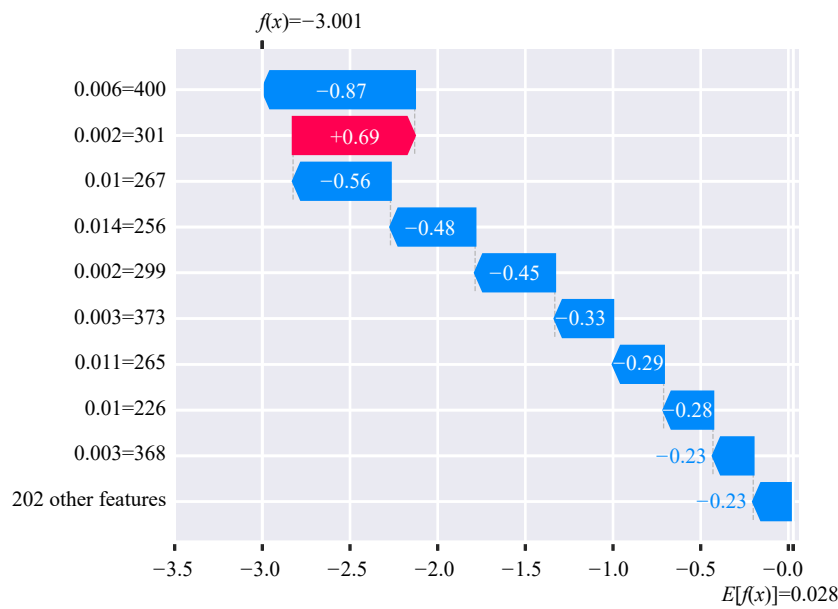


图6 薄荷寒热特征标记贡献度

效,而药效的基础是其所含物质成分,寒性中药能够调理热证,热性中药能够调理寒证,应具有相应的物质基础^[5]。在前期研究中,笔者所在课题组发现中药的化学指纹图谱可以用来识别中药的寒热药性。课

题组使用距离度量算法对中药寒热药性识别进行了尝试。由于中药紫外光谱数据大多具有样本量偏小、特征偏多等特点,导致经典分类器的识别效果一般。因此,本研究引入新的集成学习算法对中药紫外光谱

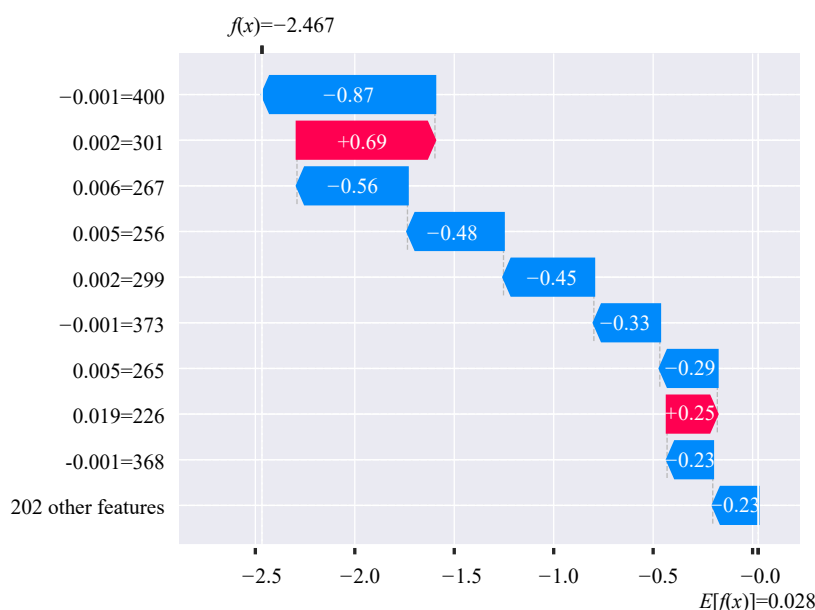


图7 车前子寒热特征标记贡献度

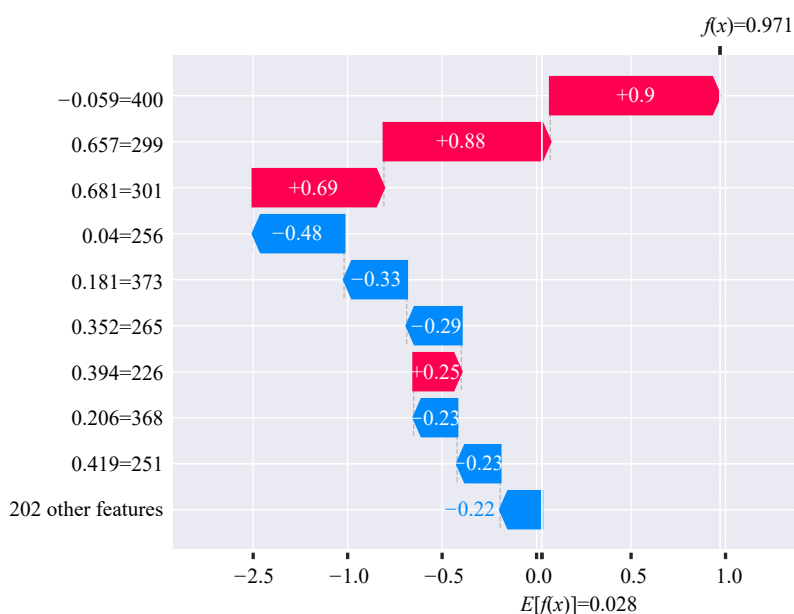


图8 白胡椒寒热特征标记贡献度

数据进行寒热药性识别。稳定性评价中,提出的模型在单溶剂下的 ACC 和 AUC 分别为 0.852、0.865;多溶剂下的 ACC 为 0.770。外推评价中,提出的模型在单溶剂下的 ACC 和 AUC 分别为 0.810、0.828;多溶剂下的 ACC 为 0.786。经过实验验证,提出的模型在单溶剂和多溶剂下均具有更好的稳定性和外推性能。实验结果表明,新的集成学习算法能够更好地适应中药紫外光谱数据的特点,与经典分类器相比,寒热药性识别效果更好。从寒热药性识别的角度,本研究初步

证明了中药的紫外光谱与中药的寒热药性存在相关性,即有相似的紫外光谱的中药具有相似的药性。初步证明了物质成分相似的中药其药性是相似的。

中药的寒热药性,寒性和热性是一个相对的概念。每一味中药都可以看作是一剂复方,无论是寒性中药还是热性中药,其内都同时含有寒性和热性物质,只是寒热物质的种类和数量的不同配伍才使中药表现为寒性和热性。当寒性物质在种类和数量上占主导地位时,该味中药表现为寒性;而当热性物质占

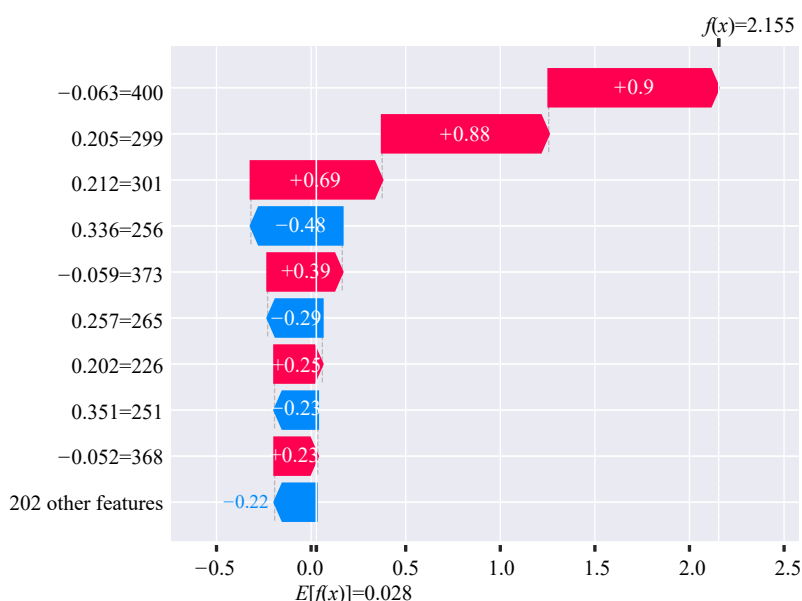


图9 芫荽寒热特征标记贡献度

主导地位时,该味中药表现为热性。本研究将一种全新的可视化方式引入中药寒热药性研究,首先对中药寒热特征标记的 feature importance 进行可视化,从而在整体层面上对决定中药寒热药性的重要特征标记进行初步的认识。然后以寒性中药薄荷、车前子,热性中药白胡椒、芫荽4种中药为例,对单味中药寒热特征标记进行可视化表达,直观地展示了各寒热特征标记对最终寒热药性识别结果的贡献度。可视化结果表明,寒性中药在紫外波长400和267 nm的吸收度具有相似性,而热性中药在紫外波长400、299和301 nm的吸收度具有相似性。从中药寒热特征标记可视化的角度,本研究进一步证明了相同药性的中药的紫外光谱具有相似性。进一步证明了如果中药的物质成分是相似的,那么药性是相似的。

4 结论

在本研究中,我们建立了一个基于集成学习的中药寒热药性识别方案。该方案通过多溶剂下的中药紫外光谱数据表征中药的物质成分,采用集成学习模型对中药紫外光谱数据进行寒热药性识别。稳定性和外推实验结果表明,本文模型能够更好地适应中药紫外光谱数据的特点,稳定性和外推性能均显著优于经典模型,在寒热药性识别上具有更高的识别准确率。本研究还通过中药寒热特征标记可视化的方式进一步分析中药寒热特征标记与寒热药性识别结果的关系。可视化的结果表明,相同药性的中药的紫外光谱具有相似性。中药寒热药性识别和寒热特征标记可视化的结果初步证明,物质成分相似的中药具有相似的药性。

参考文献

- 1 欧阳兵,王振国,王鹏. 中药四性理论现代研究思路与方法. 北京中医药大学学报, 2006, 29(9):592-594.
- 2 李睿萍,孙万森,铁晓玲,等. 探讨中药药性与温度相关性瞬时受体电位的潜在联系. 世界科学技术-中医药现代化, 2020, 22(8):2992-2997.
- 3 程薇薇,刘建利,张宁,等. 评价中药寒热药性的实验方法研究. 中草药, 2010, 41(7):1122-1126.
- 4 欧阳兵,王振国,李峰,等. 中药四性“性-效-物质三元论”假说及其论证. 山东中医药大学学报, 2008, 32(3):182-183.
- 5 王振国,王鹏,欧阳兵. 中药物质成分与寒热药性相关性研究的方法学框架. 浙江中医药大学学报, 2009, 33(5):734-737.
- 6 付先军,王鹏,王振国. 从中药“性-构关系”探索构建寒热药性成分要素表征体系的研究构想. 世界科学技术(中医药现代化), 2011, 13(5):919-924.
- 7 付先军,王振国,李学博,等. 基于化学结构相似性的植物类中药“性-构”关系分析. 中华中医药杂志, 2019, 34(6):2657-2661.
- 8 贾荣娣,汪华君,蒋俊,等. 基于动物温度趋向行为学的吴茱萸多组分寒热药性差异研究. 南京中医药大学学报, 2020, 36(1):63-67.

- 9 冯静, 李玲玲, 崔瑛. 对中药药性表征的思考. 中华中医药杂志, 2020, 35(11):5623–5625.
- 10 曹灿, 冯静, 李玲玲, 等. 基于“以效识性”观点的中药药性再认识. 中华中医药杂志, 2021, 36(2):648–653.
- 11 阙晓慧, 桂蜀华, 钟清元, 等. 不同炮制程度地黄对小鼠卵巢的影响及其寒热药性相关成分分析. 中成药, 2021, 43(11):3222–3226.
- 12 魏瑞丽, 王连心, 谢雁鸣. 中药药性认知与临床合理用药的思考. 中国中药杂志, 2021, 46(21):5462–5467.
- 13 夏元枢, 魏国辉, 王振国, 等. 中药寒热药性研究方法综述. 中华中医药杂志, 2021, 36(2):990–992.
- 14 谢欢欢, 陈晨, 王鹏. 基于科属分类下的中药物象-药性相关性研究概述. 中华中医药杂志, 2017, 32(8):3620–3622.
- 15 杨波, 王振国. 植物类中药寒热药性与无机元素相关性研究. 南京中医药大学学报, 2011, 27(2):109–111.
- 16 聂斌, 郝竹林, 桂宝, 等. 基于随机森林的中药寒、热药性代谢组学判别方法研究. 江西中医药大学学报, 2015, 27(2):82–86.
- 17 樊凤杰, 轩凤来, 白洋, 等. 基于三维荧光光谱特征的中药药性模式识别研究. 光谱学与光谱分析, 2020, 40(6):1763–1768.
- 18 Wei G H, Fu X J, Wang Z G. Multisolvent similarity measure of Chinese herbal medicine ingredients for cold-hot nature identification. *J Chem Inf Model*, 2019, 59(12):5065–5073.
- 19 魏国辉, 张丰聪, 付先军, 等. 中药成分相似性量化建模及寒热药性预测分析. 数据分析与知识发现, 2020, 4(5):75–83.
- 20 Long W, Liu P, Xiang J, *et al*. A combination system for prediction of Chinese Materia Medica properties. *Comput Methods Programs Biomed*, 2011, 101(3):253–264.
- 21 Fu X J, Mervin L H, Li X B, *et al*. Toward understanding the cold, hot, and neutral nature of Chinese medicines using in silico mode-of-action analysis. *J Chem Inf Model*, 2017, 57(3):468–483.
- 22 Wei G H, Fu X J, Wang Z G. Nature identification of Chinese herbal medicine compounds based on molecular descriptors. *J AOAC Int*, 2021, 104(6):1754–1759.
- 23 李雨, 李骁, 薛付忠, 等. 基于人工神经网络的中药药性判别研究. 山东大学学报(医学版), 2011, 49(1):57–61.
- 24 张新新, 李雨, 纪玉佳, 等. 主成分-线性判别分析在中药药性识别中的应用. 山东大学学报(医学版), 2012, 50(1):143–146.
- 25 刘文慧, 李雨, 纪玉佳, 等. 偏最小二乘在中药药性判别中的应用. 山东大学学报(医学版), 2012, 50(1):151–154.
- 26 拱健婷, 王佳宇, 李莉, 等. 基于电子鼻气味指纹图谱与XGBoost算法鉴别姜黄属中药. 中国中药杂志, 2019, 44(24):5375–5381.
- 27 Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, 2016:785–794.
- 28 Lundberg S M, Lee S I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. New York: 2017:4768–4777.
- 29 Lundberg S M, Nair B, Vavilala M S, *et al*. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*, 2018, 2(10):749–760.
- 30 Lundberg S M, Erion G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*, 2020, 2(1):56–67.
- 31 王鑫, 廖彬, 李敏, 等. 融合LightGBM与SHAP的糖尿病预测及其特征分析方法. 小型微型计算机系统, 2022, 43(9):1877–1885.

Study on Identification of Cold-Hot Nature and Visualization of Cold-Hot Characteristic Markers of Chinese Herbal Medicines Based on XGBoost and SHAP

Zhang Xike¹, Zhao Wenhua¹, Fu Xianjun², Ma Zhiqing¹, Wei Guohui^{1,2}

(1. College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan 250355, China; 2. Key Laboratory of Theory of Traditional Chinese Medicine, Ministry of Education of China, Shandong University of Traditional Chinese Medicine, Jinan 250355, China)

Abstract: Objective The XGBoost algorithm in ensemble learning was used to explore the establishment of the identification model of cold-hot nature of Chinese herbal medicines, and the relationship between cold-hot characteristic markers and the identification results of cold-hot nature of Chinese herbal medicines was further explored by SHAP. Methods Using the ultraviolet spectrum data set of 61 Chinese herbal medicines, the XGBoost algorithm in ensemble learning was used to construct the identification model of cold-hot nature of Chinese herbal medicines, the performance of the model was evaluated by stability evaluation and extrapolation evaluation, and the cold-hot characteristic markers of Chinese herbal medicines were visualized by SHAP. Results In terms of model evaluation, in the stability comparison, the ACC and AUC of this model in single solvent were 0.852 and 0.865 respectively; The ACC

under multi-solvent was 0.770. In the extrapolation comparison, the ACC and AUC of this model in single solvent were 0.810 and 0.828 respectively; The ACC under multi-solvent was 0.786. In terms of visualization, the absorption of cold Chinese herbal medicines at UV wavelengths of 400 and 267 nm was similar, while that of hot Chinese herbal medicines at UV wavelengths of 400, 299 and 301 nm was similar. Conclusion Compared with the classical model, the model in this paper has better identification ability for cold-hot nature of Chinese herbal medicines. Chinese herbal medicines with similar UV spectrum have similar nature, it is preliminarily proved that Chinese herbal medicines with similar material components have similar nature.

Keywords: Cold-Hot nature of Chinese herbal medicines, Cold-Hot characteristic markers of Chinese herbal medicines, Ensemble learning, XGBoost, SHAP, Visualization

(责任编辑: 刘玥辰, 审稿人: 王瑀、张志华)