Computer Science and Application 计算机科学与应用, 2023, 13(8), 1588-1595
Published Online August 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.138157

Hans汉斯

基于太赫兹时域光谱数据的柴胡鉴别多分类器比较

陈　帅，周楚雲，郑成勇*，刘铭蒽，张家荣，谭艳仪

五邑大学数学与计算科学学院，广东　江门

收稿日期：2023年7月17日；录用日期：2023年8月16日；发布日期：2023年8月24日

摘　要

随着机器学习领域的发展，研究人员不断探索新的分类算法模型，使得可供选择的机器学习算法种类更加丰富。然而，许多研究仅使用有限的分类算法，这导致综合比较分类器性能变得困难。为此，本实验利用柴胡太赫兹(THz)时域光谱数据，使用多个评价指标，评估了支持向量机(SVM)、KNN、决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、Logistic回归(LR)、多层感知(MLP)、伯努利朴素贝叶斯(Bernoulli Naive Bayes, BNB)、AdaBoosting、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、极端随机树(Extremely Random Forest, ERF)、极致梯度提升(eXtreme Gradient Boosting, XGB)和轻量梯度提升机(Light Gradient Boosting Machine, LGBM)等12种分类器的分类性能。结果表明，LR、MLP、SVM和KNN分类效果最好，其中，MLP的批次内投票准确率达100%，且召回率和F2得分都较为优异；此外，GBDT、AdaBoosting和LGBM等算法的柴胡鉴别准确度也普遍超过80%。本文为基于THz的柴胡鉴中的分类器选择提供了重要参考。

关键词

机器学习，分类算法，太赫兹时域光谱，柴胡

Comparison of Multiple Classifiers for Bupleurum Identification Based on Terahertz Time-Domain Spectroscopic

Shuai Chen, Chuyun Zhou, Chengyong Zheng*, Ming'en Liu, Jiarong Zhang, Yanyi Tan

School of Mathematical and Computational Sciences, Wuyi University, Jiangmen Guangdong

Received: Jul. 17th, 2023; accepted: Aug. 16th, 2023; published: Aug. 24th, 2023

*通讯作者。

文章引用: 陈帅, 周楚雲, 郑成勇, 刘铭蒽, 张家荣, 谭艳仪. 基于太赫兹时域光谱数据的柴胡鉴别多分类器比较[J]. 计算机科学与应用, 2023, 13(8): 1588-1595. DOI: 10.12677/csa.2023.138157

Abstract

With the development of machine learning, researchers are constantly exploring new classification algorithm models, making the variety of machine learning algorithms available more diverse. However, many studies only use limited classification algorithms, which makes it difficult to comprehensively compare the performance of classifiers. For this purpose, this paper used terahertz (THz) time-domain spectral data of Bupleurum to evaluate the performance of 12 classifiers including Support vector machine (SVM), KNN, Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Multilayer Perceptron (MLP), Bernoulli Naive Bayes (BNB), AdaBoosting, Gradient Boosting Decision Tree (GBDT), Extremely Random Forest (ERF), eXtreme Gradient Boosting (XGB) and Light Gradient Boosting Machine (LGBM), in terms of multiple classification performance indicators. The results showed that LR, MLP, SVM, and KNN are the four classifiers with the best classification performance. Among them, the MLP classifier reaches 100% accuracy after voting and has superior recall and F2 score; in addition, newer algorithms such as GBDT, AdaBoosting and LGBM have also been generally found to have accuracies of more than 80%. This paper provides an important reference for practical applications in the field of Chai Hu identification based on THz.

Keywords

Machine Learning, Classification, Trahertz Tme-Domain Spectral, Bupleurum

1. 引言

据报道，目前全球有柴胡属植物 200 种，我国已报道的有 43 种。尽管 1963 年版《中国药典》就规定柴胡或狭叶柴胡干燥根为柴胡正品供用药，但有研究者实际考察发现，我国药材市场流通的商品柴胡竟达十多种，多地柴胡用药不符合规定，实际应用繁乱，因此研究柴胡鉴别技术对规范柴胡市场、加强药材质量控制和促进中药产业可持续发展具有重要意义。

近年来，太赫兹光谱技术作为一项在线检测技术[1]，在农业、医学、食品安全、航天等领域应用广泛。基于太赫兹光谱技术的分类方法有很多，但如何选取一种适合数据集的分类器才是关键。基于太赫兹时域光谱数据的分类研究中使用较多的方法有 SVM、KNN 等。如在文献[2]中基于太赫兹光谱技术，结合均值偏移算法(MeanShift)和主成分分析法(PCA)，提出以支持向量机(SVM)为基础，通过改进步长和平衡全局搜索与局部搜索的策略优化布谷鸟算法(SPCS)，得到 SPCS-SVM 分类模型，提供了一种太赫兹中草药数据快速识别的方法。文献[3]中针对黄连、掺杂牛黄和天然牛黄等的太赫兹时域光谱数据，分别构建随机森林(RF)模型和三种参数优化的支持向量机(SVM)模型，对六种物质的太赫兹吸收光谱进行分类鉴别，结果表明，RF 模型和 SVM 模型均可达到 95%左右的分类准确率。文献[4]中利用三组相似中药炙甘草和生甘草、南柴胡和北柴胡、山豆根和北豆根的太赫兹光谱数据，构建三种不同的 SVM，并建立误差反向传播神经网络(BP 神经网络)，结果表明，SVM 是实现太赫兹光谱技术对中药快速、精确分类的有效方法之一。文献[5]中基于相关向量机(RVM)理论，提出了改进的多分类相关向量机(ImRVM)分类模型，实现了八种转基因棉花种子的有监督分类识别。另外，为使太赫兹光谱技术应用于鉴别时准确率更

高，测量速度更快，一些文献中引入了深度学习分类算法。如文献[6]中提出一种融合 ResNet 和长短时记忆网络(LSTM)的太赫兹时域光谱隐匿危险品识别方法，按照批次进行分析预测，以此来选择模型最优结构。文献[7]中通过对 CNN 的网络结构和重要权值参数的优化，提出了一种改进的 CNN 分类模型。该模型在提高太赫兹吸收光谱识别精度的同时，可以有效解决由于太赫兹光谱数据量不足而容易陷入局部最优的问题。以上分类算法通常具有较高准确性和稳健性，但也存在分类算法对比不足的问题。

为了解决分类算法对比不足的问题，本文提出一种多分类器测试对比方法，以太赫兹光谱柴胡数据分类为例，对 12 种分类算法进行了综合对比，以期为研究者提供有益参考。

2. 数据介绍

实验使用的柴胡太赫兹光谱数据共有 13 个批次，每个批次包含 10 个样本，总计 130 个样本，3 类柴胡。图 1 给出了所用柴胡太赫兹吸收系数谱图，其中左图为全部吸收系数谱曲线图，右图为柴胡的吸收系数谱均值曲线图，其中绿色线代表藏柴胡，蓝色线代表锥叶柴胡，红色线代表北柴胡。

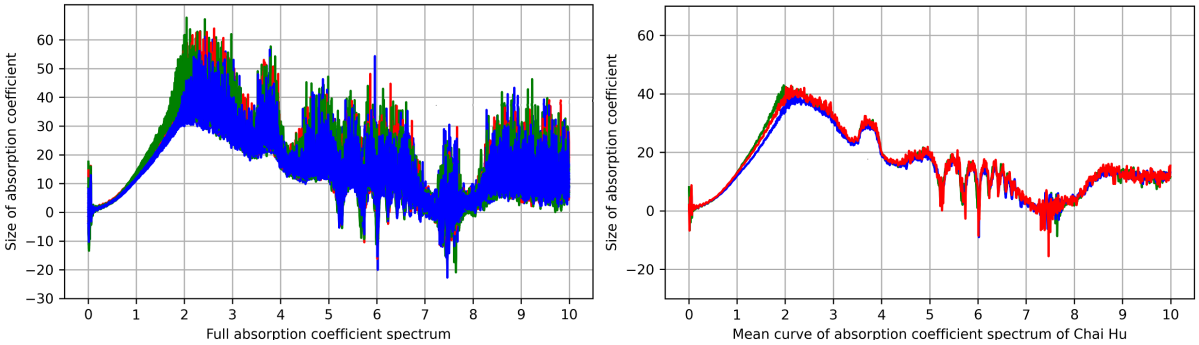

Figure 1. Spectra of the absorption coefficient of Bupleurum Terahertz
图 1. 柴胡太赫兹吸收系数光谱图

根据图 1 得，原始数据高频段噪声较多，需要对原始数据进行频段选择。经反复测试，本文选取 0.4~1.8 THz 这一相对平稳且噪声较少的频段用于后续实验，具体如图 2 所示。

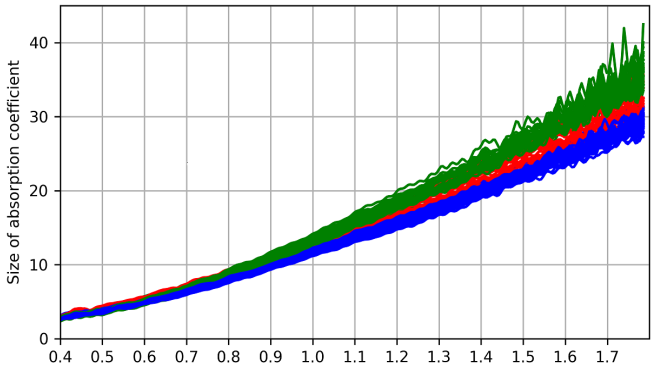

Figure 2. Terahertz absorption coefficient spectra in 0.4~1.8 THz
图 2. 0.4~1.8 THz 波段的太赫兹吸收系数光谱

为提高分类器的分类性能，本文使用正态标准化方法对数据进行预处理。该方法通过计算数据的标准差，将数据按照其与均值的偏差进行标准化处理。这种处理方式可以使得数据具有零均值和单位标准差，能有效缩放数据，减小不同频段数据之间的尺度差异，进而提高模型的性能和效率。

3. 分类器介绍

本文使用的分类器包括：支持向量机(Support Vector Machine, SVM)、最近邻(K-Nearest Neighbor, KNN)、决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、Logistic 回归(Logistic Regression, LR)、多层感知(Multilayer Perceptron, MLP)、伯努利朴素贝叶斯(Bernoulli Naive Bayes, BNB)、自适应提升(Adaptive Boosting, AB)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、极端随机树(Extremely Random Forest, ERF)、极致梯度提升(eXtreme Gradient Boosting, XGB)和轻量梯度提升机(Light Gradient Boosting Machine, LGBM)等。以下对各算法进行简要介绍：

1) SVM：是一种适用于二分类和多分类问题的有监督学习算法。其目标是找到最优超平面，以有效分开不同类别的样本并最大化分类间隔。具有丰富的核函数、通过引入软间隔和松弛变量，在一定程度上提高了 SVM 的鲁棒性和处理噪声的能力[8]。同时，SVM 对小样本集、非线性数据集表现较好。但对大规模数据，SVM 计算复杂度较高，需进行特征缩放。

2) LR：是用于二分类问题的线性回归算法。它最早由赫尔曼·菲舍尔在 20 世纪 30 年代提出，并在 50 年代由 David Cox 发展成现代逻辑回归。LR 通过逻辑函数将线性模型输出映射到概率空间，适用于线性可分和不可分问题，简单、快速且易于实现[9]。近年来，通过集成学习等方法，LR 在实际应用中变得更强大。但在处理复杂非线性问题时，LR 表现较差，容易受异常值影响。

3) KNN：是基于实例的学习算法。KNN 主要用于分类和回归，通过最近的 K 个邻居的标签进行预测。算法简单易懂，适用于非线性问题和多类别分类。但其计算复杂度较高，对样本不平衡较敏感，需要确定合适的 K 值。

4) RF：通过构建多个决策树并对其结果进行投票或平均来进行分类或回归，是一种集成学习算法。由 Leo Breiman 于 2001 年提出。如今，RF 在分类和回归问题中具有较高准确度，适用于处理大量的特征和样本，对噪声和异常值具有较好的鲁棒性。但其模型解释性相对较差，对于高维稀疏数据可能表现不佳。

5) DT：是一种基于树结构的分类和回归算法，通过逐步划分数据集，生成一棵树来做出预测，是许多集成学习算法的基础。DT 算法具有易于理解和解释的优势，不需要特征缩放，能够处理数值型和类别型特征。但 DT 容易产生过拟合，对于数据中的噪声和离群值较为敏感。

6) BNB：是朴素贝叶斯算法的一种变体，具有简单、快速的特点，适用于文本分类等特征二元分布的场景，但对于特征间相关性较强或具有连续特征的数据表现较差[10]。

7) MLP：是一种基于前馈神经网络的学习算法，由多个层次的神经元构成。随着深度学习的兴起，MLP 及其变种成为计算机视觉、自然语言处理等领域的主要算法之一。它适用于复杂的非线性问题，具有较强的拟合能力，在大规模数据集上表现优秀。但其训练时间较长，需要大量数据来避免过拟合[11]。

8) AB：是一种集成学习方法，通过加权组合多个弱分类器构建一个强分类器。作为经典集成算法，它激发了更多其他的集成学习方法的发展。AB 相对于单一弱分类器显著提高了准确率和泛化能力，尤其在高维度数据上表现优异。但对噪声和异常值相对敏感，可能导致过拟合，且计算开销较大。后续出现了许多改进算法如 GBDT 和 XGB，在解决实际问题上表现更好。

9) GBDT：是一种集成学习算法，最早由 Jerome H. Friedman 于 1999 年提出，通过迭代地训练一系列决策树并使用梯度提升策略来改进预测性能。自提出以来，其在机器学习领域广受关注，发展出了更高效的改进版本如 XGB 和 LGBM。GBDT 算法具有高准确度、较好的鲁棒性，特别擅长处理复杂问题。但需要调整大量超参数，也可能会导致过拟合问题[12]。

10) ERF：作为随机森林的改进版，致力于提高模型的泛化能力。相较于传统随机森林，ERF 引入了更多的随机性，减少过拟合风险，增强模型的多样性，有效地降低树之间的相关性，但也增加了计算开

销。在使用时，需要综合考虑优势与计算成本。

11) LGBM：是一种基于梯度提升的决策树算法[13]，旨在提高训练速度和准确度。随着大数据和工业界应用的增多，LGBM 成为梯度提升算法的重要代表，为解决大规模数据问题提供了强有力工具。该算法训练速度快，内存占用低，在大规模数据集上表现出色，但调整超参数对数据质量和噪声较为敏感。

12) XGB：在 GBDT 的基础上融合了加权策略和正则化技术，是高效的梯度提升算法[14]。XGB 具有高准确度和效率的特点，尤其适用于大规模数据集，支持并行处理和缺失值处理。但使用需要调整超参数较多，对于非结构化数据不太适合。

4. 实验设置

4.1. 批次划分方法

批次留一法适用于小规模数据，它通过按批次划分数据集来避免样本泄露导致的得分偏高的问题，使实验结果更具科学性、合理性。

结合柴胡数据多批次的特点，避免随机采样造成的样本泄露问题，本文采用按批次进行划分的方法，将数据划分为 n − 1 个批次进行分类器的训练和调试，剩余的 1 个批次数据进行测试。通过这种方式，对分类器进行 13 次训练与测试，能有效地提高了分类器的准确性和鲁棒性。

4.2. 参数设置

本文模型训练基于 Python 3.9.13 环境，采用网格搜索法对十二种分类器的参数进行调优。通过对每个分类器的各项参数可能出现的取值进行排列组合，生成一个参数组合的“网格”。遍历这个参数网格，对每个组合进行评估，并计算评估指标，如得分标准差、准确率和召回率等。通过系统地搜索所有可能的参数组合，确定最佳参数配置，从而获得给定数据集上表现最佳的分类器的最佳参数。

相关参数设置如表 1 所示，其中，C：惩罚系数，kernel：核函数，gamma：核尺度参数，n_estimators：学习器个数，min_samples_split：最少样本分割数，max_depth：学习器最大深度，inter：线性模型的截距，solver：优化器，Criterion：损失函数，weights：权重，algorithm：算法，learning_rate：学习率，min_samples_leaf：各叶子节点包含的最少样本数，Alpha：拉普拉斯或利德斯通平滑的参数，Binarize：特征二值化的阈值，class_prioc：类的先验概率，Activation：激活函数，hidden_layer_sizes：隐藏层层数及每层的节点数。

Table 1. Parameter setting for each classifier
表 1. 各分类器参数设置

Model	optimum parameter	Model	optimum parameter	Model	optimum parameter
SVM	C = 0.01 kernel = “linear” gamma = “scale”	GBDT	n_estimators = 20 min_samples_split = 100 max_depth = 3	KNN	n_neighbors = 2 weights = “uniform” algorithm = “auto”
LR	C = 1 inter = True solver = lbfgs	ERF	n_estimators = 70 criterion = “gini”	RF	n_estimators = 10 criterion = “gini”
DT	criterion = “gini” max_depth = 3 min_samples_leaf = 35	BNB	Alpha = True Binarize = True class_prioc = True	LGBM	n_estimators = 50 max_depth = 3 learning_rate = 0.01
AB	n_estimators = 30 learning_rate = 0.1	MLP	Activation = “identity” Solver = “sgd” hidden_layer_sizes = (5,5,10)	XGB	n_estimators = 10 max_depth = 3 learning_rate = 0.01

4.3. 评价准则

本实验主要的评价准则是准确率。假设测试样本正确分类的样本数量为 N_{res}，测试样本总体数量为 N，则准确率的定义为：

$$\text{Score} = \frac{N_{res}}{N}$$

由于本实验中同一批次的样本数据类别相同，故本实验基于同一批次的分类结果计算投票得分，进一步提出批次内投票前的准确率得分(Before_vote_sore)和批次内投票后的准确率得分(After_vote_score)。设投票前正确分类的样本数量为 N_{res_b}，投票后正确分类的样本数量为 N_{res_a}，则二者的计算公式为：

$$\text{Before\_vote\_sore} = \frac{N_{res\_b}}{N}$$

$$\text{After\_vote\_sore} = \frac{N_{res\_a}}{N}$$

同时，实际分类过程中，会更多的关注分类器对某一类或多类柴胡的识别性能，本实验补充召回率得分(Recall_score)。假设测试样本中属于 A 类的样本数量为 M，分类器预测为 A 类正确的样本数量为 M_{res}，则召回率的计算公式为：

$$\text{Recall\_score} = \frac{M_{res}}{M}$$

为充分考虑样本类别不均衡的情况和综合评价分类器的性能，本实验结合精确率(Precision)、召回率和标准差得分，并考虑到召回率在实际分类任务中占更重要的地位，故本文采用 F2 得分(F2_score)用于进一步完善评价准则，假设分类器预测为 A 类的样本数量为 H，其中预测正确的样本数量为 H_{res}，则 F2 的计算公式为：

$$\text{Precision} = \frac{H_{res}}{H}$$

$$\text{F2\_score} = 5 * \frac{\text{Precision} * \text{Recall\_score}}{4 * \text{Precision} + \text{Recall\_score}}$$

为展示各度量的稳定性，本实验引入标准差指标。标准差是一组数据平均值分散程度的一种度量。标准差数值越大，代表大部分数值与其平均值之间差异较大；标准差越小，代表数值接近平均值，总体比较稳定。假设样本数据的算数平均值为，样本数据的数量为 n，则标准差的计算公式为：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

5. 实验结果及分析

本实验采用多种分类模型对样本数据进行分类[15]，在实验前需要对样本数据进行批次划分，随后进行数据标准化处理。各种分类模型在最优参数下的分类情况如表 2、图 3 所示，score1、score2、score3、score4 分别对应 Before_vote_score、After_vote_score、Recall_score、F2_score。其中，score1、score2 通过加减标准差显示其得分稳定性。

Table 2. Scores of each classifier
表 2. 各分类器得分

model_name	score1	score2	score3	score4	Use_time
SVM	**93.85 ± 3.47**	92.31 ± 7.10	0.89	0.90	0.14
LR	**94.36 ± 1.77**	**100.00 ± 0.00**	0.89	0.91	0.19
KNN	93.08 ± 4.52	92.31 ± 7.10	0.89	0.90	**0.07**
RF	80.00 ± 9.67	84.62 ± 13.02	0.63	0.65	0.10
DT	82.31 ± 12.75	84.62 ± 13.02	0.76	0.76	0.09
GBDT	83.85 ± 10.24	84.62 ± 13.02	0.77	0.78	3.47
ERF	79.23 ± 8.85	84.62 ± 13.02	0.66	0.68	0.18
AB	83.08 ± 10.06	84.62 ± 13.02	0.72	0.73	2.05
LGBM	80.00 ± 12.77	84.62 ± 13.02	0.74	0.75	1.19
BNB	73.85 ± 8.85	84.62 ± 13.02	0.60	0.63	0.10
MLP	**96.92 ± 0.67**	**100.00 ± 0.00**	**0.91**	**0.91**	0.91
XGB	78.46 ± 11.67	84.62 ± 13.02	0.62	0.64	0.46

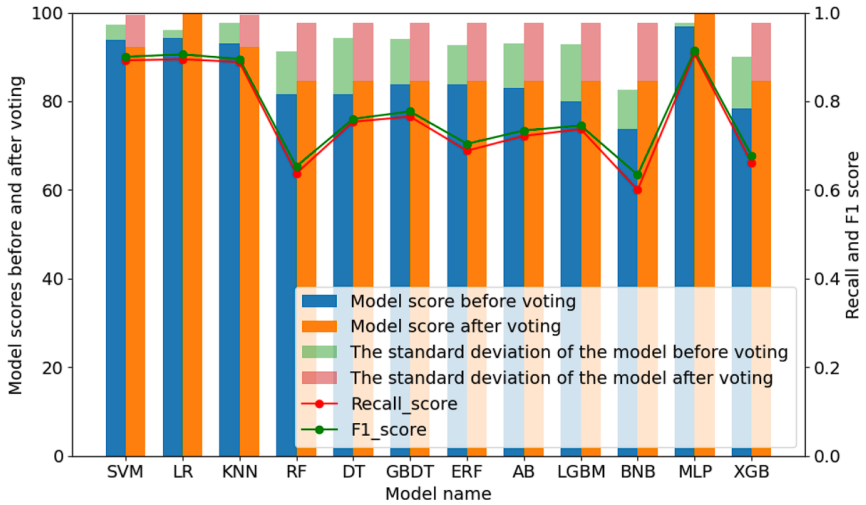

Figure 3. Accuracy of different classifiers before and after voting
图 3. 投票前后不同分类器的准确率

由表 2 可知，各模型分类精度普遍较好，在投票后，各模型都能达到 80%以上的准确率。对比分析表 2 和图 3 可知，投票前准确率得分最佳的模型是 MLP，得分 96.92、其次是 LR、SVM，得分最低的是 BNB；投票后准确率得分最佳的分类模型是 MLP、LR，得分 100，其次是 SVM、KNN，其余分类器得分一致；而用时最短的分类模型是 KNN，最长的是 GBDT。综合各衡量指标，MLP 和 LR 分类模型性能最优，其中 MLP 在重要指标上都优于 LR 模型。因此，对本样本数据而言，最优的分类模型是 MLP，其次是 LR，最差的是 BNB。

由于 THz 数据为连续特征数据，BNB 分类器不适用于处理连续特征，导致在所有分类器中，它的分类效果最差。相比之下，MLP 算法由大量的非线性神经元组成，拥有较强的拟合能力和解决非线性问题的能力，因此在所有分类器中，它的分类效果最好。

6. 结束语

本实验通过可视化太赫兹光谱柴胡数据，截取频谱范围在 0.4~1.8 THz 内的数据，根据批次划分训练集和测试集，使用 12 种不同的分类器模型对数据进行训练与测试并输出得分结果。结合得分标准差、准确率和召回率进行综合性评估分析得出结论：在众多分类模型中，传统的 MLP、LR、SVM 及 KNN 表现较优，优于当前一些热门的分类方法如 GBDT、AB、LGBM 等；在众多文献中表现优异的 ERF、XGB 表现欠佳。本文实验结果表明，没有一流的算法，只有合适的算法。

本文的相关结果对研究人员在后续进行太赫兹光谱数据分类时具有一定的借鉴意义。仍有部分问题需进一步的研究与探索：

(1) 本文使用的批次留一法具有很好的避免样本泄露问题，但该方法仅适合小规模数据。而在使用大规模数据集时，则存在实验次数过多的问题，需采用其他按批次的随机样本划分方法。

(2) 本文使用网格搜索进行参数调优对于参数较少时效果较好，当参数较多时，搜索空间急剧增大，导致搜索效率低下。在后续的研究中，可考虑其他如遗传算法、贝叶斯优化方法等，以提高搜索效率。

(3) 本文的 THz 频段选择方法主要基于直观和经验，后续可进一步探索排序和搜索的频段特征选择方法，以进一步提高算法性能。

参考文献

[1] 杨惠智, 杨婷, 孙万阳, 郭萍, 孙国祥, 李茜, 李晓辉. 中药一致性评价新技术——中药太赫兹光谱发展及其量子指纹图谱在中药一致性评价中的应用[J]. 中南药学, 2022, 20(7): 1478-1486.

[2] 盘书宝. 基于太赫兹光谱的中草药快速识别及含量检测方法研究[D]: [博士学位论文]. 桂林: 桂林电子科技大学, 2022.

[3] 章龙. 基于太赫兹光谱技术与化学计量学方法的中药识别研究[D]: [硕士学位论文]. 南京: 南京林业大学, 2020.

[4] 陈艳江, 刘艳艳, 赵国忠, 等. 基于支持向量机的中药太赫兹光谱鉴别[J]. 光谱学与光谱分析, 2009, 29(9): 2346-2350.

[5] 庹帅. 基于太赫兹时域光谱技术的转基因农产品种子识别方法研究[D]: [硕士学位论文]. 武汉: 武汉科技大学, 2022. DOI:10.27380/d.cnki.gwkju.2022.000591

[6] 赵聪. 融合 ResNet 和 LSTM 的太赫兹时域光谱数据识别方法[J]. 工业控制计算机, 2022, 35(9): 90-92.

[7] 郑志杰, 林振衡, 谢海鹤, 等. 基于卷积神经网络的工程塑料太赫兹光谱分类识别方法[J]. 光谱学与光谱分析, 2023, 43(5): 1387-1393.

[8] 杨超宇, 陈雯君, 耿显亚. 基于改进 SVM 的中文专利文本分类比较研究[J]. 武汉理工大学学报(信息与管理工程版), 2023, 45(2): 292-298+303.

[9] 吉黎明, 熊兴旺, 杨子荣. 一种基于逻辑回归的柴油机工况分类模型[J]. 小型内燃机与车辆技术, 2023, 52(2): 6-9+20.

[10] 张丽娟, 夏艳, 程雪平, 等. 基于伯努利贝叶斯模型的高校贫困生预测研究[J]. 信息技术与信息化, 2021(11): 159-161.

[11] 谢永康, 丁梦清, 徐啸, 等. 基于 MLP 神经网络算法的中医肥胖体质分类模型研究[J]. 无线互联科技, 2021, 18(7): 37-40.

[12] 刘鸿浩, 杨玲玲. 基于 GBDT 算法的多因子选股策略研究[J]. 产业创新研究, 2023(9): 124-126.

[13] 何芸. 基于 LGBM 模型的城市道路交通流量预测研究[J]. 电子技术与软件工程, 2022(3): 259-262.

[14] 甘思雨. 基于 XGBoost 算法的多因子选股策略研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2022.

[15] 刘畅畅. 数据分类算法性能的大规模实验对比分析[D]: [硕士学位论文]. 郑州: 河南大学, 2016.