第40卷，第9期
2020年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 40，No. 9，pp2696-2701
September，2020

太赫兹光谱技术在生物活性肽检测中应用研究

王 璞[1]，何明霞[1*]，李 萌[2]，曲秋红[2]，刘 锐[3]，陈永德[4]

1. 天津大学测试计量技术及仪器国家重点实验室，天津 300072
2. 莱仪特太赫兹(天津)科技有限公司，天津 300019
3. 天津科技大学食品工程与生物技术学院，天津 300222
4. 百德福生物科技有限公司，河北 唐山 063000

摘 要 生物活性肽作为21世纪人类健康的新宠儿，研究证明其对人体生命活动有着很好的作用，其检测方法也是备受关注，太赫兹时域光谱技术因为其独特的性质在检测生物活性肽中有着不可比拟的优势。选用牛骨肽、海参肽和牛肽这三种生物活性肽，通过透射式太赫兹时域光谱系统得到其在0.5～2 THz的吸收系数曲线。从太赫兹吸收系数曲线来看，鱼肽吸收系数大于海参肽和牛骨肽。因为生物活性肽的氨基酸种类和肽键的相互作用，导致其在太赫兹频段内没有明显的吸收峰，为了更好的对其进行检测区分，建立分类判别模型，寻找出最适合这类物质的方法。在对太赫兹原始吸收系数数据进行S-G平滑处理，归一化预处理之后，随机选取四分之三预处理好的数据划分为训练集，其余为预测集，导入分类判别模型。模型包括分类器和最优参数选取两部分，分类器选取支持向量机，随机森林和极限学习机等有监督的分类方法，使用遗传算法、粒子群算法和网格搜索等智能优化算法选取支持向量机最优参数。为了减少原始光谱数据维数并提高模型的运算速度，使用主成分分析进行预处理，将降维之后的结果导入分类模型。综合考虑其准确率和运行时间等因素，虽然基于粒子群算法的支持向量机具有最高的准确率98.3%，但是运行时间较长为180 s；使用极限学习机能够有着最短的运行时间0.2 s，但是准确率为73.3%。基于网格搜索的支持向量机准确率为95%，运行时间为11 s，能够在准确率较高的情况下使用较短的时间，证明基于网格搜索的支持向量机对生物活性肽太赫兹吸收光谱具有快速，准确的分类结果。研究结果表明，利用太赫兹时域光谱技术结合机器学习算法能够实现快速、无损检测生物活性肽，为生物活性肽的检测提供了一种新思路，同时也为THz-TDS结合机器学习对吸收峰不明显的多肽之间的鉴别提供参考。

关键词 生物活性肽；太赫兹时域光谱技术；机器学习；主成分分析

中图分类号：O436 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)09-2696-06

引 言

生物活性肽是一类分子介于蛋白质和氨基酸之间，由多种氨基酸以一定方式结合而成的二肽到多肽，具有一定生理作用的低分子聚合物[1]。生物活性肽相比于单个氨基酸，更容易且更有效被人体吸收，适合于年老体弱，过敏体质的人群。相比于蛋白质生物大分子，能够发挥其整体结构所不具有的特殊功能。具有降低血压，抗衰老，促进消化吸收及提高自身免疫调节能力等作用。在功能食品，药品，疫苗制备等食品学和医学领域有着广泛的应用。因此对它们的检测一直是国内外学者研究的重点。目前国内外主要应用的分析方法为色谱法，质谱法，核磁共振光谱[2]。

太赫兹(Terahertz，THz)辐射是指波长在0.03～3 mm之间，频率在0～10 THz，介于红外和微波之内的光谱[3]。THz波具有很好的透过性和特征光谱性质，运用其特性可以进行物质非接触式鉴别。多肽有其特定的氨基酸组成，且相互之间有电偶极矩，使其易受到太赫兹波段作用。Kutteruf[4]等通过改变温度，得到固相短肽链的THz吸收光谱吸收峰变化，又通过改变肽链氨基酸的数量，发现其吸收系数曲线变得复杂。文献[5]报道了四种简单二肽的太赫兹吸收曲线和各自在0～2.7 THz的吸收峰，并且通过对双甘氨肽、

收稿日期：2019-07-28，修订日期：2019-11-05

基金项目：国家自然科学基金项目(61675151)资助

作者简介：王 璞，1997年生，天津大学精密仪器与光电子工程学院硕士研究生 e-mail：wp412005676@163.com

*通讯联系人 e-mail：hhmmxx@tju.edu.cn

丙谷二肽、肌肽和谷胱甘肽这四种肽分子结构的分析和密度泛函理论模拟，认为肽键的差异会导致肽类分子对太赫兹的吸收产生差别。

对于无明显太赫兹吸收峰的物质，一般难以通过吸收系数谱进行分类识别，需要结合机器学习算法和化学计量法进一步进行处理。通过建立有效的分析模型与太赫兹光谱技术相结合将是这个方面的重点内容。选择的预测模型为有监督的学习算法，包括支持向量机[6](support vector machine，SVM)，随机森林[7](random forest，RF)，极限学习机[8](extreme learning machine，ELM)。支持向量机的主要思想是结构风险最小化的近似实现。但是由于支持向量机会由于数据维数过大而分类拟合效果不好等问题，本文结合主成分分析进行降维比较。随机森林是一种根据统计的思想，根据决策树的判断类别得出结果的分类器，拥有高预测精度和运算量小等特点。极限学习机是一种针对传统单隐前馈神经网络而提出的分类模型，有学习速度快，泛化性能好等优点。为了提高预测速度，降低噪声干扰，选择主成分分析法[9]进行对比，主成分分析法(principal component analysis，PCA)是一种常用的可以用于降维的方法，能够在丢失较少特征信息的前提下，将较高维度的数据转化为较低维度的数据。为了能寻找到支持向量机中参数的最优值，选择网格搜索(grid search，GS)，粒子群优化(particle swarm optimization，PSO)和遗传(genetic algorithm，GA)[10]算法作为优化算法。其中网格搜索通过穷举搜索选取最优参数；遗传算法通过一系列内在机制，仿照种群的进化过程，得到适应度近似最优的状态；粒子群算法不断调整速度和位置参数，来寻求最优解。

本文主要利用海参肽、牛骨肽、鱼肽三种代表性生物活性肽的太赫兹光谱数据，结合不同的机器学习算法，创建分类模型。主要以测试集预测准确率为考察标准，以运行速度为辅助标准。通过太赫兹光谱技术结合机器学习分类方法在生物活性肽检测领域进行探索。

1　实验部分

1.1　设备

实验使用的是日本 advantest 公司的 TAS 7500SU。光谱范围为 0.5～7.0 THz，动态范围为 57 dB，频率分辨率为 7.6 GHz。本实验中用的是其透射模块，其结构如图 1 所示。

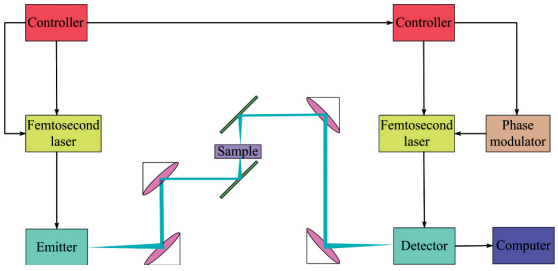

图 1　太赫兹时域光谱系统

Fig. 1　Schematic of THz-TDS

1.2　样品制备

实验中所用的牛骨肽粉末，海参肽粉末，鱼肽粉末均由百德福生物科技有限公司提供，纯度为 99%，白色粉末。实验中为了保证测量的稳定性，将样品在压片之前置于干燥柜中干燥 6 h，干燥柜湿度 20%，温度 30 ℃。将样品与聚乙烯按照 2∶1 的质量比例混合，充分研磨。在 10 MPa 压力下，压 5 min，压成厚度为(1.1±0.1) mm，直径为 13 mm 的样品片，每种多肽分别压制符合要求，表面均匀的样品各 30 片。

1.3　方法

在实验中，以干燥空气作为参考信号，每片样品分别在不同的位置测量 3 次。为了保证结果的可重复性和精确性，样品测完第一次之后放入干燥柜中保存 24 h，进行复测，同样也是每片样品移动不同位置分别测量 3 次。得到每片样品的吸收系数谱。

1.4　数据处理方法

为了降低系统噪声和实验因素导致的噪声，提高光谱的平滑性，使用 Savitzky-Golay(S-G)平滑预处理，考虑原光谱的特性，将平滑滤波器的拟合阶数设置为 3 阶，并且考虑其平滑特性，设置每 15 个点平滑一次。由于光谱图两端噪声比较大，选取 0.5～2 THz 范围内的 198 个光谱数据进行分析。将数据进行标准化处理，归一化到[0，1]范围内。如图 2 所示，使用主成分分析法，光谱数据降维到 8 维之后的贡献率之和为 95%，可以代替原光谱图。

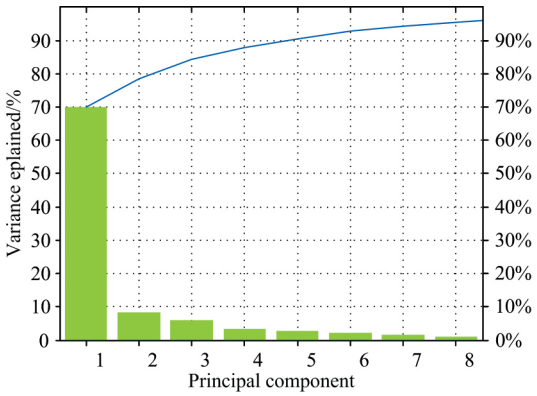

图 2　PCA 各成分得分

Fig. 2　PCA component score

分类模型如图 3 所示，其中对于支持向量机参数优化环节，选择网格搜索、遗传算法和粒子群算法对其参数优化。训练模型选择的是支持向量机、随机森林和极限学习机。结果主要考察分类准确度和运行时间，在确保准确率高，大于 90%的前提下，考虑运行时间。

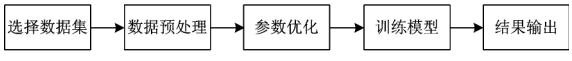

图 3　模型流程图

Fig. 3　Model flow chart

2　结果与讨论

2.1　吸收系数谱

将两次测量得到的数据进行平均，为了表示三种多肽的

不确定度大小，三种多肽在 0.5～2 THz 范围内的误差棒如图 4 所示，从图中可以看出在低频段，三种多肽样品几乎重叠，难以直接区分；在高频段，区分度较好，鱼肽吸收系数明显大于海参肽和牛骨肽。从这些多肽的太赫兹吸收系数上不能很容易对其进行区分，需要采用一些机器学习的算法。

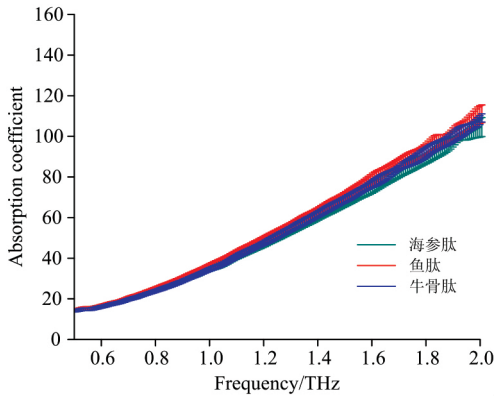

图 4 海参肽、鱼肽和牛骨肽的吸收系数误差棒

Fig. 4 Absorption coefficient error bars for sea cucumber peptides, fish peptides and bovine bone peptides

2.2 建模及定性分析

将经过数据预处理后的全部样品加上标签，随机选取四分之三数据量进行算法训练，其余数据用来进行测试。

为了找到分类三种多肽最好的算法，采用不同的机器学习方法进行对比验证，图 5(a)为网格搜索加 5 折交叉验证法的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 8，核函数参数 g 的最优值是 0.125，训练集准确率 81.1%，测试集准确率 95%，用时 11.7 s。

图 5(b)为遗传算法寻优加 5 折交叉验证法的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 0.79，核函数参数 g 的最优值是 356.3，训练集准确率 63.9%，测试集准确率 85%，用时 152.8 s。

图 5(c)为粒子群寻优加 5 折交叉验证法的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 83.44，核函数参数 g 的最优值是 0.01，训练集准确率 82.2%，测试集准确率 98.3%，用时 180.8 s。

图 5(d)为主成分分析结合网格搜索下的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 1.414，核函数参数 g 的最优值是 2，训练集准确率 73.3%，测试集准确率 78.3%，用时 6.27 s。

图 5(e)为主成分分析结合遗传算法下的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 1.543，核函数参数 g 的最优值是 2.2，训练集准确率 81.7%，测试集准确率 78.3%，用时 41.9 s。

图 5(f)为主成分分析结合粒子群算法下的支持向量机模型结果，结果表明，向量机惩罚因子 C 的最优值为 1.5，核函数参数 g 的最优值是 1.7，训练集准确率 82.2%，测试集准确率 75%，用时 65.3 s。

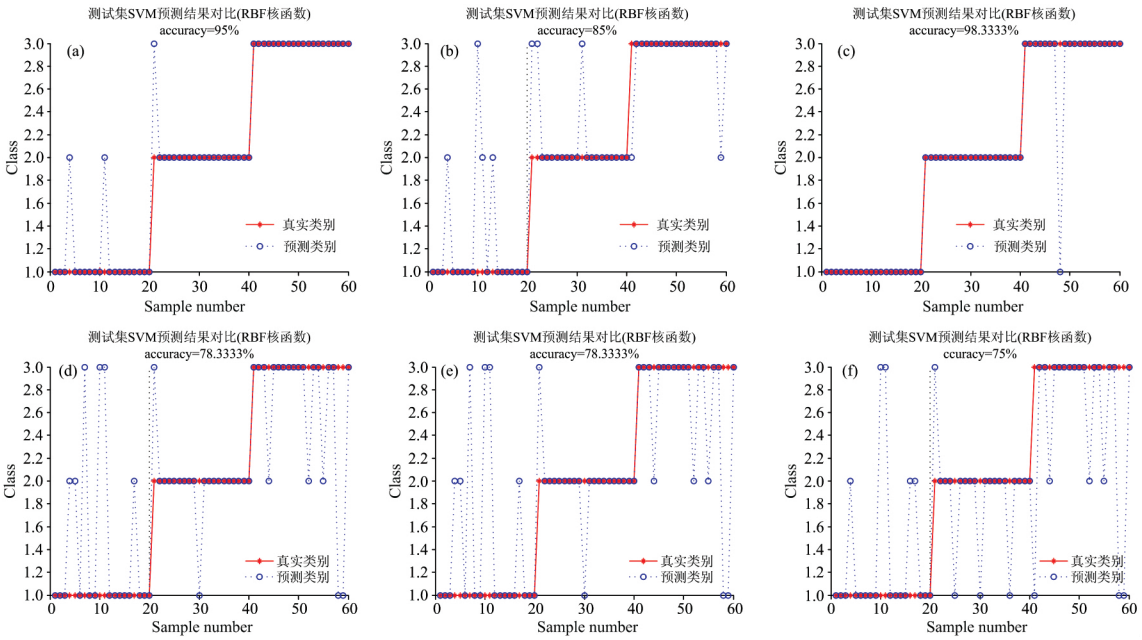

图 5 不同优化方法下支持向量机分类结果

(a)：GS-SVM；(b)：GA-SVM；(c)：PSO-SVM；(d)：PCA-GS-SVM；(e)：PCA-GA-SVM；(f)：PCA-PSO-SVM

Fig. 5 SVM classification results under different optimization methods

(a)：GS-SVM；(b)：GA-SVM；(c)：PSO-SVM；(d)：PCA-GS-SVM；(e)：PCA-GA-SVM；(f)：PCA-PSO-SVM

建立随机森林模型，经过多次试验，综合考虑准确率和运行时间，参数选择如图 6(a)所示，最优的决策树个数为 400，准确率达到最优准确率，时间最短。随机森林模型本身自带降维的能力，无需进行降维处理，结果如图 6(b)所示，

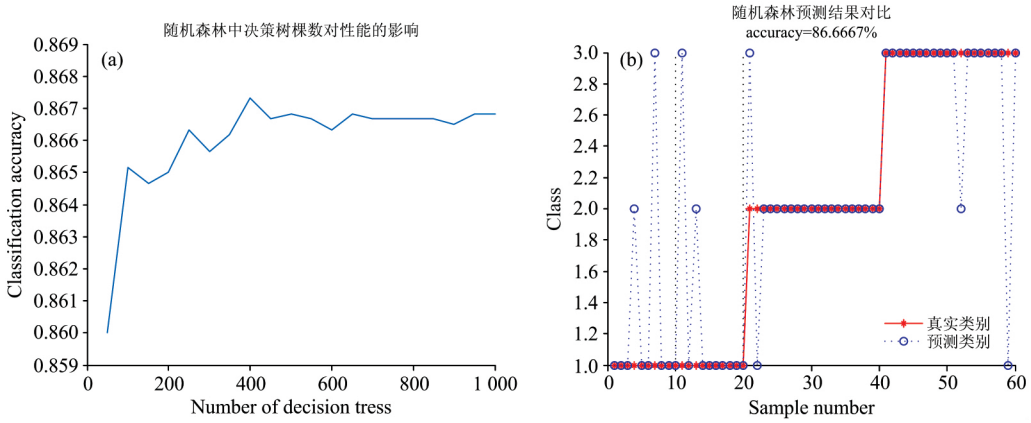

图 6　随机森林分类结果

(a)：最优参数选择；(b)：RF 分类结果

Fig. 6　The classification result of RF

(a)：Optimal parameter selection；(b)：RF classification results

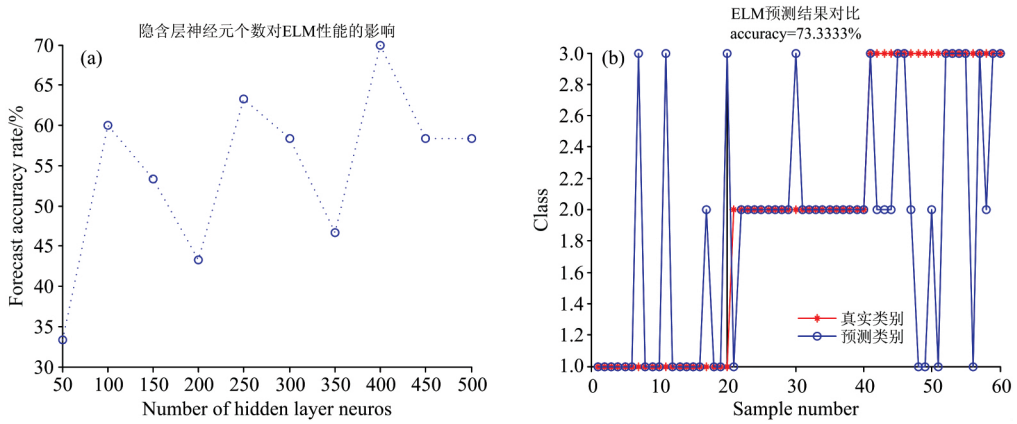

图 7　ELM 分类分类结果

(a)：最优参数选择；(b)：ELM 分类结果

Fig. 7　The classification result of ELM

(a)：Optimal parameter selection；(b)：ELM calssification results

准确率为 86.6%。

建立极限学习机模型，经过多次试验，综合考虑准确率和运行时间，参数选择如图 7(a)所示，最优的隐含层神经元个数为 400，准确率达到最高。极限学习机结果如图 7(b)所示，准确率为 73.3%。

表 1 给出了多种分类方法的预测精度和运行时间。从表 1 看出，数据进行 PCA 预处理之后，测试集的准确率较未进行预处理有所下降，但是运行时间也加快。通过比较三种监督机器学习算法，准确率最高的是支持向量机，但是运行时间最快的是极限学习机。准确率最高的是基于粒子群算法的支持向量机分类，为 98.8%(59/60)的准确率。运行时间最短的是极限学习机，只需要 0.2 s。但是，在综合考虑测试集准确率和运行时间的情况下，最适合分类这三种多肽的算法是基于网格搜索的支持向量机，准确率为 95%(57/60)，运行时间是 11.7 s。

表 1　建模方法对预测结果的影响

Table 1　The impact of modeling methods on forecasting results

分类方法	预处理	训练集准确率/%	测试集准确率/%	时间/s
GS-SVM	—	81.1	95	11.7
	PCA	73.3	78.3	6.27
GA-SVM	—	63.9	85	152.8
	PCA	81.7	78.3	41.9
PSO-SVM	—	82.2	98.3	180.8
	PCA	82.2	75	65.3
RF	—	90.4	86.7	1.83
ELM	—	89.4	73.3	0.20

4　结　论

以牛骨肽，海参肽，鱼肽三种生物活性肽为研究对象，

验证了太赫兹时域光谱技术对其定性分析中的应有潜力。为了更好的对其进行区分，利用这些多肽的吸收光谱信息结合机器学习算法，并且比较数据在 PCA 降维之后和未降维的分类对比情况，得出最适合分类这些多肽的分类算法。结果证明，使用网格搜索的支持向量机结合太赫兹时域光谱技术，可以实现对多肽的高效鉴别，有望促进太赫兹时域光谱技术在生物医学检测领域的应用。

致谢：感谢百德福生物科技有限公司对样品的支持，以及莱仪特太赫兹(天津)科技有限公司提供太赫兹系统。

References

[1] ZHANG Zhi-hui, SU Xiu-lan(张志慧，苏秀兰). China Medical Herald(中国医药导报), 2019, 16(10): 37.
[2] ZHOU Ting-yi, GAO Xin-chang, et al(周亭屹，高新昌，等). Science and Technology of Food Industry(食品工业科技), 2019, (12): 335.
[3] HE Ming-xia, GUO Shuai(何明霞，郭 帅). Journal of Electronic Measurement and Instrument(电子测量与仪器学报), 2012, 26(8): 663.
[4] Kutteruf M R, Brown C M, Iwaki L K, et al. Chemical Physics Letters, 2003, 375(3/4): 337.
[5] LI Li-long(李利龙). Master Degree Dissertation(硕士学位论文). Changsha University of Science and Technology(长沙理工大学), 2014.
[6] HE Xiao-qun(何晓群). Multivariate Statistical Analysis(多元统计分析). Beijing: China Renmin University Press(北京：中国人民大学出版社), 2008. 152.
[7] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
[8] Olaru C, Wehenkel L. Fuzzy Sets and Systems, 2003, 138(2): 221.
[9] Breiman L. Machine Learning, 2001, 45(1): 5.
[10] Bendu H, Deepak B B V L, Murugan S. Applied Energy, 2017, 187: 601.

Application of Terahertz Spectroscopy in the Detection of Bioactive Peptides

WANG Pu[1], HE Ming-xia[1*], LI Meng[2], QU Qiu-hong[2], LIU Rui[3], CHEN Yong-de[4]
1. State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China
2. LET Terahertz (Tianjin) Technology Co., Ltd., Tianjin 300019, China
3. School of Food Engineering and Biotechnology, Tianjin University of Science and Technology, Tianjin 300222, China
4. Bai Defu Biological Technology Co., Ltd., Tangshan 063000, China

Abstract Bioactive peptides, as the new darling of human health in the 21st century, have been proved that they have a good effect on human life activities, and their detection methods are also of great concern. Terahertz time-domain spectroscopy technology has incomparable advantage in detecting bioactive peptides because of its unique properties. In this paper, three bioactive peptides, bovine bone peptide, sea cucumber peptide and fish peptide, were used to obtain the absorption coefficient curve of 0.5～2 THz by the transmission terahertz time domain spectroscopy system. From the terahertz absorption coefficient curve, the absorption coefficient of the fish peptide is higher than that of sea cucumber peptide and fish bone peptide. Because of the interaction between the amino acid species of bioactive peptides and peptide bonds, there is no obvious absorption peak in the terahertz frequency band. In order to better detect and distinguish them, a classification discriminant model is established to find the most suitable for such substances. After the S-G smoothing and normalization preprocess performed on the terahertz original absorption coefficient data, two-thirds of the pre-processed data are randomly selected into training sets, and the rest are prediction set. The classification discriminant model is introduced. The model includes two parts: the classifier and the optimal parameter selection. The classifier selects the supervised classification method such as support vector machine, random forest and extreme learning machine, and uses the intelligent optimization algorithm such as genetic algorithm, particle swarm optimization and grid search to select the support vector machine optimal parameters. In order to reduce the original spectral data dimension and improve the computational speed of the model, Principal Component Analysis is used for preprocessing, and the results after dimensionality reduction are imported into the classification model. Considering the factors such as accuracy and running time, although the support vector machine based on particle swarm optimization has the highest accuracy rate of 98.3%, the running time is longer than 180 seconds; the ultimate learning machine can have the shortest running time of 0.2 seconds.

However, the accuracy rate is 73.3%. The support vector machine based on grid search has an accuracy rate of 96% and a running time of 11 seconds. It can use a shorter time in the case of higher accuracy, and proves that the support vector machine based on grid search is better for detecting bioactive peptide. The results show that the use of terahertz time-domain spectroscopy combined with machine learning algorithms can achieve rapid and non-destructive detection of bioactive peptides, providing a new idea for the detection of bioactive peptides. It also demonstrates that THz-TDS combined with machine learning is a way better way for the identification of inconspicuous peptides.

Keywords Bioactive peptide; Terahertz time-domain spectrum; Machine learning; PCA

* Corresponding author

(Received Jul. 28, 2019; accepted Nov. 5, 2019)