基于太赫兹时域光谱和模式识别技术软玉和仿品鉴别

林红梅[1]，曹秋红[1]，张同军[1]，李照鑫[1]，黄海青[1]，
李学敏[1]，吴 斌[2]，张庆建[3]，吕新民[4]，李德华[1*]

1. 山东科技大学电子信息工程学院，青岛市太赫兹重点实验室，山东 青岛 266590
2. 中国电子科技集团公司第四十一研究所，山东 青岛 266555
3. 青岛海关技术中心，山东 青岛 266002
4. 阿拉山口海关技术中心，新疆 阿拉山口 833400

摘 要 玉石是一种稀有的矿物质，自古以来备受国人喜爱，其真伪鉴别一直是珠宝鉴别行业的棘手难题，传统的鉴别方法已经难以实现对真假玉石的准确鉴别。太赫兹检测技术可以实现快速无损检测，在混合物的分类鉴别方面，有广泛的应用。基于太赫兹时域光谱技术和模式识别技术，对来自我国新疆、青海，以及巴基斯坦、阿富汗四个地区的软玉样品及玻璃、大理石、石包玉三种仿品，使用透射模式测得样品在0.1～1.5 THz频率范围内的太赫兹谱，通过参数提取得到其折射率谱线。由于其化学成分的复杂和多样性，仅靠其特征谱线图，并不能正确的区分软玉和仿品，为了更好的对其进行鉴别，需要建立分类模型。采用主成分分析(PCA)对实验得到的原始折射率数据进行降维和特征提取，作出样品在第一、二主成分上的二维得分图，在图中可以看出软玉和仿品能够很明显的区分开来。在经过降维处理之后的数据中，随机选取其中的四分之三作为训练集，剩下的作为测试集，输入到支持向量机(SVM)建立的分类模型中，并引入网格搜索(GridSearch)、遗传算法(GA)和粒子群算法(PSO)对支持向量机参数进行优化。结果显示，基于网格搜索的支持向量机最优参数 $c=2.828\ 4$，$g=2$，识别率为97.7%，运行时间为1.39 s，用时最短；基于遗传算法的支持向量机最优参数 $c=1.740\ 1$，$g=4.544\ 6$，识别率为98.3%，运行时间为3.6 s；基于粒子群算法的支持向量机最优参数 $c=11.287\ 2$，$g=1.833\ 1$，识别率为98.6%，运行时间为6.13 s，用时最长。虽然三种优化算法得到的最优参数不同，但均可实现正确的分类。研究结果表明，使用太赫兹时域光谱技术结合模式识别方法可以快速、准确的鉴别软玉和仿品，这为玉石的鉴别提供了一种新手段。

关键词 软玉；太赫兹时域光谱；主成分分析；支持向量机

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)11-3352-05

引 言

玉石有软玉、硬玉之分，平常人们所说的玉多指软玉，而硬玉指的是翡翠。玉与石的主要区别就是玉的质地较为细腻，富有韧性，呈半透明状，且有光泽；而石基本上是没有光泽的，且入手粗糙，通常是不透明的。随着加工技术的进步，玉石仿品的做工可以以假乱真，单靠肉眼很难鉴别。因此很多现代科技手段被用于玉石鉴别。例如红外光谱技术，但是该技术需要已知的样品光谱参数，并且光谱分析工作难度较大；拉曼光谱技术[1]中荧光现象会造成很大的背景干扰，且进行傅里叶变换时，常出现曲线的非线性问题。因此寻找一种实用、便捷、准确可靠的玉石无损检测技术极为重要。

由于太赫兹波对非金属材料具有很好的穿透性，光子能量低、使用安全，且具有很宽的波谱范围，因此被广泛用于无损检测和安检成像。孟倩等[2]使用太赫兹时域光谱技术对玉石和仿品进行分析，试图根据其折射率、吸收系数以及介电常数的差别来鉴别和田玉的真伪。杨婷婷[3]等使用太赫兹时域光谱技术对不同产地的白色软玉进行研究，根据光谱折射率的数值差异，以及特征吸收峰的不同来区分不同产地的软玉。但是大部分的软玉在太赫兹波段没有特征吸收峰，只根据其特征谱的差异，不能准确的对软玉进行鉴别。

收稿日期：2020-10-26，**修订日期**：2021-03-29

基金项目：国家重点研发计划“变革性技术关键科学问题”重点专项(2017YFA0701000)，国家重点研发计划项目(2018YFF0215400)资助

作者简介：林红梅，1995年生，山东科技大学电子信息工程学院硕士研究生 e-mail：1664741597@qq.com

*通讯作者 e-mail：jcbwl@sdust.edu.cn

利用太赫兹时域光谱技术结合模式识别方法对软玉和仿品进行鉴别。实验测量软玉和仿品的折射率，使用主成分分析(principal component analysis，PCA)对原始折射率数据进行降维处理。通过支持向量机(support vector machines，SVM)建立相应的分类模型，并引入网格搜索(Grid Search)、遗传算法(genetic algorithm，GA)和粒子群算法(particle swarm algorithm，PSO)对 SVM 的相关参数进行优化，实现了对软玉和仿品的有效识别。

1　实验部分

1.1　装置

本实验中使用的太赫兹时域光谱系统是由德国 BATOP 公司生产的 TDS-1008，仪器光路示意图如图 1 所示。实验在恒温、恒湿下进行。本实验采用透射模式测量样品太赫兹时域谱。

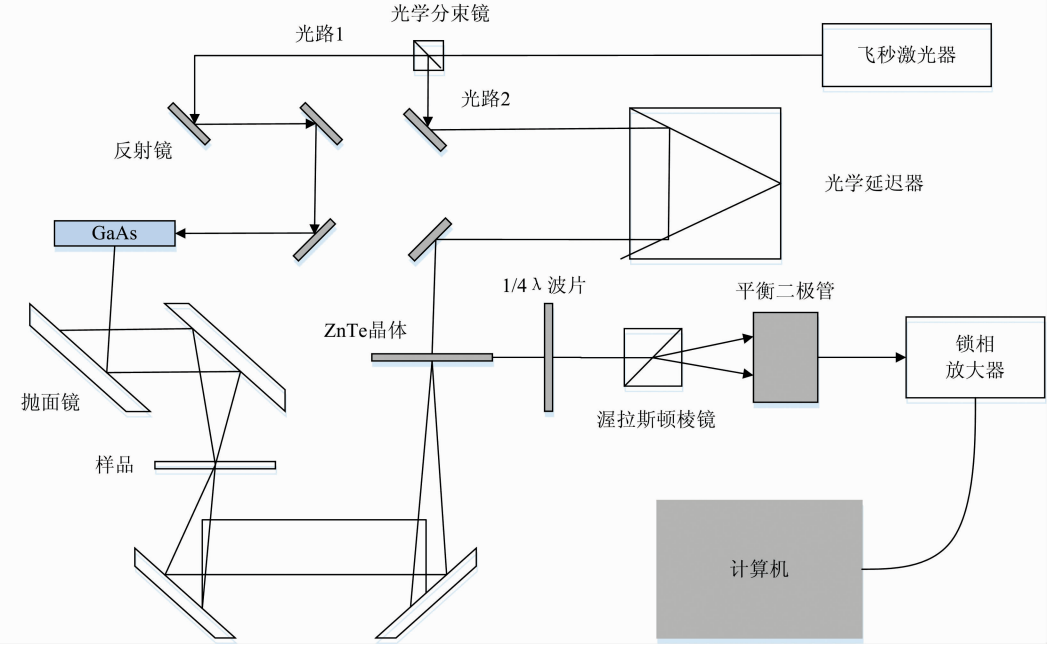

图 1　THz-TDS 实验原理图

Fig. 1　Experimental schematic diagram of THz-TDS

1.2　样品

实验选用来自我国新疆、青海，以及巴基斯坦、阿富汗四个地区的软玉样品，仿品选用玻璃、大理石、石包玉三种样品，样品表面光滑，厚度在 3 mm 左右。使用太赫兹时域光谱系统测得样品的折射率，其有效光谱范围为 0.1～1.5 THz。每个地区软玉样品各测得 12 组数据，四个地区共 48 组数据，仿品共测得 12 组数据，软玉和仿品数据共 60 组。

1.3　方法[4]

主成分分析(PCA)是一种统计方法，该方法通过正交变换把高维的原始数据空间映射到一个小维度的空间，即通过提取包含原始数据信息的特征数据(主成分)，组成一个新的低维数据集[5-6]。主成分 PC1 包含原始数据信息最多，其次是主成分 PC2，主成分 PC3，…，且各主成分两两正交。求解主成分的步骤如下。

(1)对原始数据矩阵 $\boldsymbol{X}_{n\times p}$($n$ 为样本的数量，p 为数据的维度)进行标准化

$$X_i^* = \frac{X_i - \mathrm{mean}(X_i)}{\mathrm{std}(X_i)} \quad i = 1,2,\cdots,p \tag{1}$$

(2)计算样本的相关系数矩阵 $\boldsymbol{R}_{p\times p}$；

(3)计算样本相关系数矩阵 $\boldsymbol{R}_{p\times p}$的特征值 λ_i 和相应的特征向量 μ_i；

(4)提取重要主成分，一般而言，当前 k 个主成分的累计方差贡献率超过 85%时，就可以用前 k 个主成分代替原始数据。

1.4　支持向量机[7-8]

支持向量机是一种分类方法，它的基本思想是寻找一个能够把特征数据准确无误的分割开，且具有最大几何间距的分离超平面。超平面的表达式如(2)所示

$$f(x) = \omega^{\mathrm{T}}x + b \tag{2}$$

式(2)中：x 为折射率光谱数据经降维后提取出来的特征向量；ω 和 b 分别表示超平面的法向量及对应的截距。

求解最优超平面，就要使两类样本之间的间距达到最大，即 L 达到最小，L 的表达式为

$$L = \frac{1}{2}\|\omega\|^2 \tag{3}$$

为了能够将全部的数据点正确分类在超平面的两侧，L 需满足

$$L = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(\omega^{\mathrm{T}}x_i + b) - 1] \tag{4}$$

式(4)中：α 为拉格朗日乘子，$\beta \geqslant 0$；x_i 为要分类的数据点；y_i 为根据映射函数得到的值。当数据线性不可区分时，就需要将其映射到一个高维空间，把数据转换成线性可分再进行

分类。通过引入核函数来避免数据在高维空间计算困难。在此选择径向基函数作为核函数。核函数 $K(x_i,x_j)$可表示为

$$K(x_i,x_j)=\exp(-|x_i-x_j|^2/\sigma^2) \quad (5)$$

通过核函数映射后，L 的表达式可转化为

$$L=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{n}\alpha_i[y_i(\omega^{\mathrm{T}}K(x_i)+b)-1] \quad (6)$$

只要确定了式(6)中的 ω 和 b，即可得到最优超平面。

2 结果与讨论

2.1 光谱分析

使用 MATLAB 软件分别对我国新疆、青海，以及巴基斯坦、阿富汗四个地区软玉样品和玻璃、大理石和石包玉三种仿品的时域光谱进行傅里叶变换处理，得到每种样品的频域谱，如图 2(a)所示。由于样品对太赫兹波有一定的吸收，因此样品的光谱振幅会有所降低。图 2(b)是样品的折射率谱。从图 2 可以看出，无法通过特征谱线区分软玉和仿品。

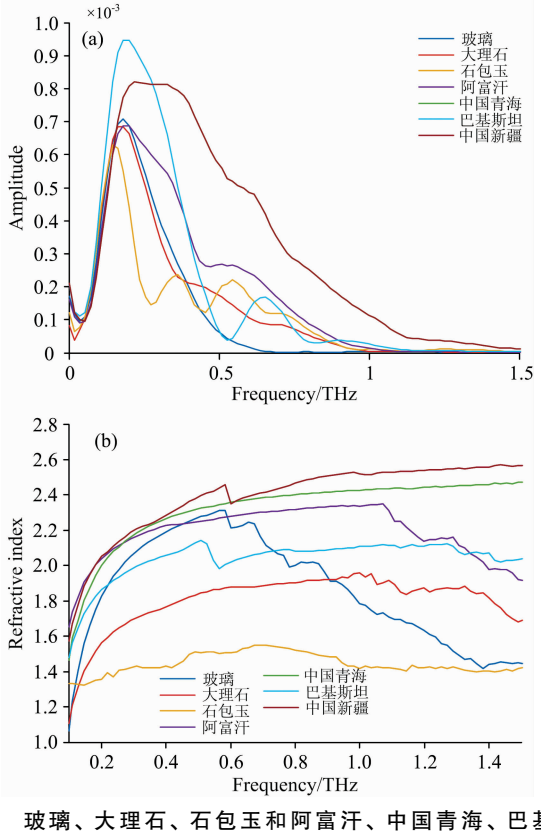

图 2 玻璃、大理石、石包玉和阿富汗、中国青海、巴基斯坦、中国新疆四个地区玉石样品的太赫兹(a) 频域谱，(b) 折射率

Fig. 2 (a) Terahertz frequency spectrum, (b) refractive index of glass, marble, raw gemstone and Jades from Afghanistan, China's Qinghai, Pakistan and China's Xinjiang

2.2 主成分分析

为了去除光谱中的重叠信息以及与样品性质不相关的信息，缩短模型的计算时间、提高运行效率[9]，将提取的 0.1～1.5 THz 频率范围内折射率 60×78 的原始数据减少到 60×4(选取方差累计贡献率最高的 4 个主成分)，折射率的各主成分的方差贡献率以及累计方差贡献率如表 1 所示，前四个主成分的总贡献率高达 98.408%，因此前四个主成分被认为在很大程度上代表了原始折射率谱的光谱特征。图 3 为样品在第一、二主成分上的得分，从图中可以看出，软玉和仿品可以很明显的区分开来，不同地区的软玉也有聚合现象，但几种软玉聚合相对比较集中，所以此种方法对于不同地区的软玉无法进行区分。

表 1 折射率各主成分方差贡献率及累计方差贡献率

Table 1 Variance contribution index and cumulative variance contribution index of each principal component of refractive index

成分	方差贡献率/%	累计方差贡献率/%
PC1	66.192	66.192
PC2	26.861	93.053
PC3	3.844	96.897
PC4	1.511	98.408

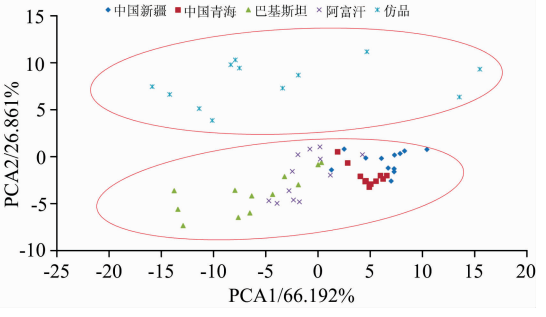

图 3 阿富汗、中国青海、巴基斯坦、中国新疆四个地区玉石样品和仿品在第一、二主成分上的得分

Fig. 3 Scores of the first and second principal components of jade samples from Afghanistan, China's Qinghai, Pakistan and China's Xinjiang and imitations

2.3 支持向量机分析

在进行主成分分析后，用新数据矩阵(60×4)代替原来的光谱数据矩阵并输入到 SVM 中建立分类模型。在 SVM 中，数据集被分为两类，一类作为训练集，一类作为测试集。随机抽取包含软玉和仿品在内的 45 组数据作为训练集，剩下的 15 组数据作为测试集。

分类模型的性能主要取决于惩罚参数 c 和径向基函数核参数 g 的选择。为了达到期望的分类效果，模型参数的选择尤为重要，因此分别采用网格搜索法、遗传算法、粒子群算法[10]对参数进行优化。

首先选用网格搜索法对参数 c 和 g 进行优化，建立网格搜索-支持向量机模型，图 4 为网格搜索选择 SVM 参数的结果。

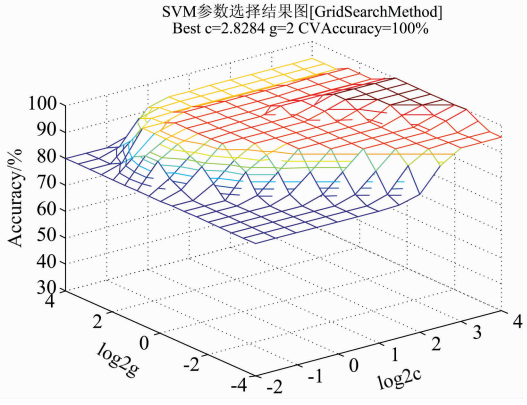

图 4　网格搜索-支持向量机参数选择结果
（最优参数 $c=2.828\ 4$，$g=2$）

Fig. 4　Result of GridSearch-SVM parameter selection
(optimal parameter $c=2.828\ 4$, $g=2$)

遗传算法的灵感来自于连续几代生物遗传特性的变化和生物的自然选择，该算法通过迭代从群体中选取较优的个体[9]。这里将 GA 的相关参数进行如下设置：最大进化代数设为 200、种群数量设为 20、将 c 的范围设定在(0～100)之间、将 g 的范围设定在(0～1 000)之间、交叉验证数设为 5，其仿真结果如图 5 所示。从图中可以看出利用遗传算法寻找出的最优参数($c=1.740\ 1$，$g=4.544\ 6$)可以使训练集分类准确率达到 100%，算法的平均适应度约为 97%。

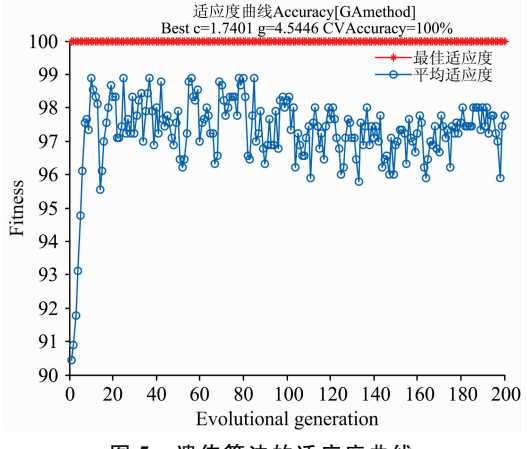

图 5　遗传算法的适应度曲线
（最优参数 $c=1.740\ 1$，$g=4.544\ 6$）

Fig. 5　Fitness curve of GA
(optimal parameter $c=1.740\ 1$, $g=4.544\ 6$)

粒子群优化算法的灵感来自于动物群体之间的社会互动。它首先用一组粒子表示一个可能的优化方案，然后通过迭代搜索最优解[11]。这里将 PSO 的相关参数进行如下设置：学习因子 C1 代表局部搜索能力设为 1.5、C2 代表全局搜索能力设为 1.7、进化代数设为 200、种群数设为 10、将 c 的范围设定在(0.1～100)之间、将 g 的范围设定在(0.01～1 000)之间、交叉验证数设为 5，其仿真结果如图 6 所示。从图中可以看出利用粒子群算法寻找出的最优参数($c=11.287\ 2$，$g=1.833\ 1$)可以使训练集分类准确率达到 100%，算法的平均适应度约为 86%。

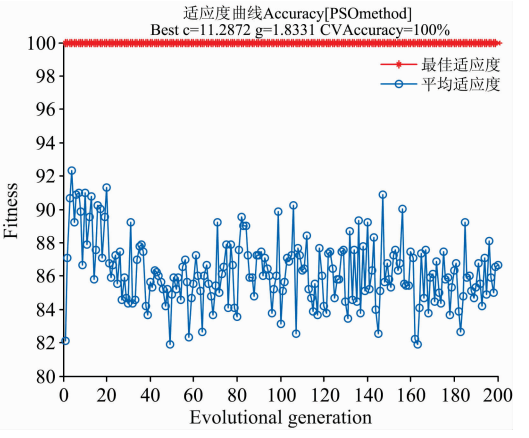

图 6　粒子群算法的适应度曲线
（最优参数 $c=11.287\ 2$，$g=1.833\ 1$）

Fig. 6　Fitness curve of PSO
(optimal parameter $c=11.287\ 2$, $g=1.833\ 1$)

将三种支持向量机参数优化方法进行对比，相关参数如表 2 所示，其中分类准确率为 20 次分类的平均值。从表中可以看出这 3 种优化方法均可以获取分类器的最优参数，虽然参数并不相同但基本可以实现正确分类，识别率分别为 97.7%，98.3%和 98.6%。

表 2　支持向量机结合网格搜索、遗传和粒子群三种优化方法对比

Table 2　Comparison of three optimization methods of SVM combined with Gridsearch, GA and PSO

优化方法	最优参数 c	最优参数 g	种群数量	迭代次数	建模耗时/s	测试集准确率/%
网格搜索法	2.828 4	2	20	200	1.39	97.7
遗传算法	1.740 1	4.544 6	20	200	3.60	98.3
粒子群算法	11.287 2	1.833 1	20	200	6.13	98.6

3　结　论

将太赫兹时域光谱技术与支持向量机相结合，建立了软玉和仿品的分类器。采用主成分分析对原始折射率数据进行降维和特征提取，将提取后的结果输入到支持向量机建立的模型中。引入网格搜索法、遗传算法和粒子群算法对支持向量机参数进行优化。三种算法的优化识别率分别为 97.7%，98.3%和 98.6%，实验结果表明，太赫兹时域光谱结合支持向量机模型能够实现软玉和仿品的有效识别。这种通过太赫兹时域光谱技术结合模式识别的方法，为真假软玉的鉴别提供了一种新的方法。

References

[1] WANG huan, WANG Yong-zhi, ZHAO Yu, et al(王 欢, 王永志, 赵 瑜, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(7): 2050.

[2] MENG Qian, BAO Ri-ma, YANG Qing-ning, et al(孟 倩, 宝日玛, 杨清宁, 等). Modern Scientific Instruments(现代科学仪器), 2015, (2): 87.

[3] YANG Ting-ting, WANG Xuan, HUANG Bo, et al(杨婷婷, 王 璇, 黄 博, 等). Acta Petrologica et Mineralogica(岩石矿物学杂志), 2020, 39(3): 314.

[4] LIAN Fei-yu, FU Mai-xia, GE Hong-yi, et al(廉飞宇, 付麦霞, 葛宏义, 等). Chinese Oils and Fats(中国油脂), 2017, 42(7): 69.

[5] Schweizer K, Cattin P C, Brunner R, et al. J. Biomech., 2012, 45(13): 2306.

[6] Noori R, Sabahi R, Karbasst A R, et al. Desalination, 2010, 260(1/3): 129.

[7] LIU Jun-xiu, DU Bin, DENG Yu-qiang, et al(刘俊秀, 杜 彬, 邓玉强, 等). Chinese Journal of Lasers(中国激光), 2019, 49(6): 0614039.

[8] 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.

[9] MA Yong-jie, YUN Wen-xia(马永杰, 云文霞). Application Research of Computers(计算机应用研究), 2012, 29(4): 1201.

[10] Bendu H, Deepak B B V L, Murugan S. Applied Energy, 2017, 187: 601.

[11] Liang J, Guo Q-J, Chang T-Y, et al. Optik, 2018, 174: 7.

Identification of Nephrite and Imitations Based on Terahertz Time-Domain Spectroscopy and Pattern Recognition

LIN Hong-mei[1], CAO Qiu-hong[1], ZHANG Tong-jun[1], LI Zhao-xin[1], HUANG Hai-qing[1], LI Xue-min[1], WU Bin[2], ZHANG Qing-jian[3], LÜ Xin-min[4], LI De-hua[1*]

1. Qingdao Key Laboratory of Terahertz Technology, College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China
2. The 41st Research Institute of CETC, Qingdao 266555, China
3. Technology Center of Qingdao Customs, Qingdao 266002, China
4. Technology Center of Alashankou Customs, Alashankou 833400, China

Abstract Jade is a rare mineral that people have favored. The identification of jade authenticity has always been a thorny problem in the jewelry identification industry. Traditional identification methods are difficult to identify the nephrite and their imitations. Terahertz standoff detection technology can realize quick non-destructive testing and has a variety of applications in the classification and identification of mixtures. In this paper, Terahertz Time-domain Spectroscopy (TDS) and pattern recognition are applied to identify nephrite and imitations. The terahertz spectrum of several nephrite jade samples from Afghanistan, China's Qinghai, Pakistan and China's Xinjiang and imitations, like glass, marble, and raw gemstone is measured with TDS in the frequency range 0.1～1.5 THz. Due to the complexity and diversity of the sample's chemical composition, the nephrite jade and the imitation cannot be distinguished correctly withtheir characteristic spectrum. In order to distinguish Jade with their imitations, a classification model is established. Principal Component Analysis (PCA) performs dimension reduction and feature extraction on the refractive index. The scores of the first and second principal components of the sample were obtained. It can be found that nephrite and imitations can be clearly distinguished from each other. Based on the extracted data, third quarters of them are randomly selected as the training set, the rest as the test set, a Support Vector Machine (SVM) model is established, and the parameters of the Support Vector Machine is optimized by GridSearch, genetic algorithm (GA) and particle swarm algorithm (PSO). The optimal parameters of SVM based on grid search are $c=2.828\ 4$ and $g=2$ while that based on GA are $c=1.740\ 1$, $g=4.544\ 6$ and based on PSO $c=11.287\ 2$, $g=1.833\ 1$. The recognition rates of the three optimization algorithms are 97.7%, 98.3% and 98.6%, and the running time is 1.39, 3.6, 6.13 s respectively. Although the optimal parameters obtained by the three optimization algorithms are different from each other, all of them can achieve a correct classification. The results show that the Terahertz spectrum combined with the pattern recognition method is a promising technique for identifying nephrite with their imitations.

Keywords Nephrite; Terahertz time-domain spectrum; Principal component analysis; Support vector machine

* Corresponding author

(Received Oct. 26, 2020; accepted Mar. 29, 2021)