文章编号：2095-4980(2020)02-0190-07

基于多特征联合的太赫兹药品检测方法

王天鹤[a]，吴紫阳[a]，丁金闪[*a]，张玉洪[b]

(西安电子科技大学 a.雷达信号处理国家重点实验室；b.电子工程学院，陕西 西安 710071)

摘　要：以药品为研究对象，利用太赫兹时域光谱系统对 3 种不同药品进行测量并提取折射率、介电常数和物质因子等多个特征参数，然后联合多个特征参数作为输入，采用后向传播(BP)神经网络、支持向量机(SVM)和学习矢量化(LVQ)3 种机器学习方法分别对药品进行多特征联合检测分类识别。实验结果表明，多特征联合检测方法识别准确率能够达到 95%以上，有效提高药品的检测分类准确度，可用于药品的检测和分类识别。

关键词：太赫兹时域光谱技术；多特征联合；机器学习；药品检测

中图分类号：TN29　　**文献标志码：**A　　**doi：**10.11805/TKYDA2019038

A multi-feature joint medicine inspection method based on THz-TDS

WANG Tianhe[a]，WU Ziyang[a]，DING Jinshan[*a]，ZHANG Yuhong[b]

(a.National Key Lab of Radar Signal Processing；b.School of Electronic Engineering，Xidian University，Xi'an Shaanxi 710071，China)

Abstract：Based on the Terahertz Time-Domain Spectrum(THz-TDS) technology, a multi-feature joint medicine inspection method is proposed to study three different medicines. First, the measurement data is acquired with a THz-TDS system. For consistency, medicines are dried, crushed into powder and then made into capsules. Then, different features like refractive index and material factor are obtained by using feature extraction method. Finally, medicines are classified and identified by a multi-feature joint detection method, in which three different machine learning methods, Back Propagation(BP) neural network, Support Vector Machine(SVM) and Learning Vectorization Quantization(LVQ), are adopted to improve the efficiency and accuracy. In the training process, all parameters are combined as the training sample in order to improve the characteristic ability. Experiment results show that the accuracy with machine learning are above 95%, and for SVM, the accuracy reaches 99%. The results confirm the application of the terahertz multi-feature joint method in medicine quality inspection and identification.

Keywords：Terahertz Time-Domain Spectrum technology；multi-feature joint；machine learning；medicine inspection

近年来，随着高效稳定的太赫兹发射源和高灵敏度的太赫兹探测器的研制成功，太赫兹技术成为研究热点[1]。太赫兹时域光谱(THz-TDS)技术可在较宽的频带内同时获得材料透射或反射前后的电磁波幅度和相位变化，且太赫兹波波长和生物的大分子尺寸接近，对很多物质太赫兹波探测可获得更丰富的特征信息，因此该技术在材料、医学、安检和探测成像等领域具有广泛应用[2-5]。

太赫兹光谱技术在药物化学领域的研究多集中在对药物种类或成分的定性鉴别上[6]。张琪等利用太赫兹技术对抗结核药物进行了定量定性分析[7]；王婷等利用太赫兹技术开展了抗生素类药物的识别研究[8]；刘乔等利用太赫兹技术研究了 3 种阿莫西林胶囊的识别[9]；肖春阳等研究了奶粉中山梨酸钾的太赫兹光谱特性[10]；刘英等利用太赫兹光谱技术研究了食品添加剂的检测分析[11]。这些研究证实了太赫兹时域光谱技术在药品及食品的质量监控和分类识别中有广阔的应用前景。

近年来，机器学习得到了飞速发展，实现了大量成功应用[12-13]。后向传播(BP)神经网络算法是一种按误差逆传播算法训练的多层前馈网络，是目前最成功的神经网络算法之一。支持向量机(SVM)是建立在统计学习理论

收稿日期：2019-01-26；**修回日期：**2019-03-13

作者简介：王天鹤(1992–)，男，在读博士研究生，研究方向为 SAR 成像、太赫兹成像。email:wangtianhe1992@126.com

[*]**通信作者：**丁金闪　email:ding@xidian.edu.cn

的 VC(Vapnik and Cortes)维理论和结构风险最小原理基础上的有监督学习方法，在解决小样本、非线性及高维模式识别中表现出许多特有优势。学习向量量化(LVQ)是一种基于自组织神经网络的有监督学习算法，是一种结构简单、功能强大的神经网络分类方法。本文运用 THz-TDS 系统测定了 3 种不同药物的光谱信息，通过计算其吸收光谱、折射率、飞行时间等多个特征，采用 BP,SVM 和 LVQ 3 种方法进行多特征联合的药品检测，准确率可达 95%以上，可用于药品的检测识别。

1　THz-TDS 系统原理

THz-TDS 系统原理如图 1 所示。THz-TDS 系统为宽频带太赫兹辐射源，其重复频率为 100 MHz，系统中最大动态范围超过 70 dB，频谱范围大于 3 THz。该系统使用外部飞秒激光器作为信号源，激光波长 1 560 nm，激光分为 2 路：一路发送到太赫兹发射器，由光电导天线产生太赫兹辐射脉冲，通过 2 个透镜聚焦在图 1 所示的样品上，然后通过 2 个透镜汇聚到太赫兹接收器，转换为电信号；另一路激光信号通过光学扫描延迟线到达太赫兹探测器，2 路信号在探测器经过处理再经过放大后输入计算机，经过傅里叶变换处理得到载有被测爆炸物信息的太赫兹时域脉冲的波形、振幅和相位等参数。

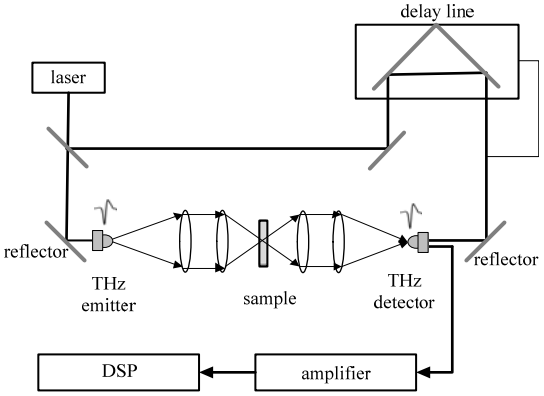

Fig.1 Block diagram of the THz-TDS system
图 1 THz-TDS 系统原理图

在数据录取中，太赫兹波透过空气的信号为参考信号，透过药品后的信号为样品信号，通过数字信号处理可以提取药品的多种特征信息，然后进一步对药品进行分类识别。

2　太赫兹多特征提取方法

太赫兹时域光谱具有指纹谱特性，包含被测物品丰富的频谱信息，因此本文通过提取折射率、吸收率、介电常数及物质因子等多种特征参数对探测物品进行分类识别。

图 2 为太赫兹波透过样品时的路径图，太赫兹波垂直入射，$E_{\mathrm{THz}}(t)$ 为发射信号，没有样品时接收到的信号作为参考信号 $E_{\mathrm{ref}}(t)$，放置样品时的信号作为样品信号 $E_{\mathrm{sam}}(t)$，对 2 个信号做傅里叶变换得到 $E_{\mathrm{ref}}(\omega)$ 和 $E_{\mathrm{sam}}(\omega)$，则传输函数为：

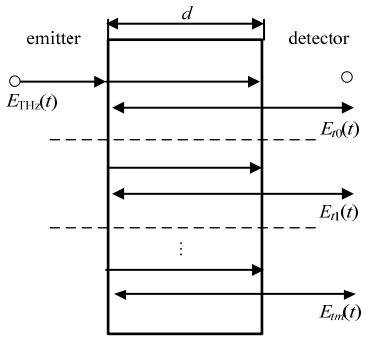

Fig.2 Terahertz wave path through sample
图 2 太赫兹波入射及反射路径

$$H(\omega)=\frac{E_{\mathrm{sam}}(\omega)}{E_{\mathrm{ref}}(\omega)}=\frac{4\tilde{n}}{(1+\tilde{n})^2}\exp\left(\frac{-\mathrm{j}d[\tilde{n}(\omega)-1]}{c}\right)FP(\omega) \tag{1}$$

$$FP(\omega)=\sum_{m=0}^{\infty}\left(\frac{1+\tilde{n}}{1-\tilde{n}}\right)^{2m}\exp\left[-\mathrm{j}(2m)\tilde{n}(\omega)d\omega\right] \tag{2}$$

式中：$FP(\omega)$ 为多次反射而叠加引起的变化因子；m 为第 m 次反射，如图 2 中的 $E_{tm}(t)$ 所示，实验中样品厚度较大，二次回波与一次回波间距较大，通过时域分离可使 $FP(\omega)$ 近似为 1；$\tilde{n}(\omega)$ 为复折射率；d 为样品厚度；空气的折射系数为 1。将复传输函数表示为模和辐角形式，可得

$$H(\omega)=\rho(\omega)\exp[-\mathrm{j}\phi(\omega)] \tag{3}$$

式中：$\rho(\omega)$ 为传递函数幅度；$\phi(\omega)$ 为相位。将复折射率 $\tilde{n}(\omega)=n(\omega)-\mathrm{j}\kappa(\omega)$ 代入式(1)，得

$$\rho(\omega)=\frac{4[n^2(\omega)+\kappa^2(\omega)]^{1/2}}{[n^2(\omega)+1]^2+\kappa^2(\omega)}\exp\left[-\frac{\kappa(\omega)d\omega}{c}\right] \tag{4}$$

$$\phi(\omega)=\frac{[n(\omega)-1]\omega d}{c}+\arctan\left[\frac{\kappa(\omega)}{n^2(\omega)+n(\omega)+\kappa^2(\omega)}\right] \tag{5}$$

若满足 $n(\omega)\gg\kappa(\omega)$ 条件，则 $\arctan\left[\dfrac{\kappa(\omega)}{n^2(\omega)+n(\omega)+\kappa^2(\omega)}\right]$ 近似为 0，可得到折射率 $n(\omega)$、消光系数 $\kappa(\omega)$、吸收率 $\alpha(\omega)$ 和介电常数 $\varepsilon(\omega)$ 分别为：

$$\begin{cases} n(\omega)=\dfrac{\phi(\omega)c}{\omega d}+1 \\ \kappa(\omega)=\dfrac{c}{\omega d}\ln\left(\dfrac{4n(\omega)}{\rho(\omega)(n(\omega)+1)^2}\right) \\ \alpha(\omega)=\dfrac{2\omega\kappa(\omega)}{c}=\dfrac{2}{d}\ln\left(\dfrac{4n(\omega)}{\rho(\omega)(n(\omega)+1)^2}\right) \\ \varepsilon(\omega)=(n(\omega)-\mathrm{j}\kappa(\omega))^2 \end{cases} \tag{6}$$

考虑到多个参数受到样品厚度的影响，采用迭代法进一步提高参数的精确度。定义传递函数误差为：

$$Err=\sum_{\omega}\left(\left\|H_{\text{theory}}(\omega)\right|-\left|H_{\text{real}}(\omega)\right\|+\left|\angle H_{\text{theory}}(\omega)-\angle H_{\text{real}}(\omega)\right|\right) \tag{7}$$

式中：$H_{\text{theory}}(\omega)$ 表示理论上的传递函数；$H_{\text{real}}(\omega)$ 表示实际测量得到的传递函数。Duvillaret[14]等证明，厚度越准确，传递函数误差越小。因此，首先估计药品的厚度区间，然后将估计的厚度区间划分为多个子区间，分别计算各个厚度对应的误差，取最小误差的厚度为准确厚度。然后以该厚度为中心取一个小的区间，再划分区间进行计算。经过多次迭代可以得到最终的厚度，从而计算出药品更准确的特征参数。

为去除药品包装带来的影响，对参数进一步进行处理。对 2 个厚度为 d_1, d_2 的药品测量得到 2 组参数：

$$\begin{cases} n(\omega)-1=\dfrac{\phi_1(\omega)c}{\omega d_1},\ \ \kappa(\omega)=\dfrac{c}{\omega d_1}\ln\left(\dfrac{4n(\omega)}{\rho_1(\omega)(n(\omega)+1)^2}\right) \\ n(\omega)-1=\dfrac{\phi_2(\omega)c}{\omega d_2},\ \ \kappa(\omega)=\dfrac{c}{\omega d_2}\ln\left(\dfrac{4n(\omega)}{\rho_2(\omega)(n(\omega)+1)^2}\right) \end{cases} \tag{8}$$

整理后消去厚度参数可得：

$$Q=\frac{\kappa(\omega)}{n(\omega)-1}=\frac{\ln\left(\dfrac{\rho_2(\omega)}{\rho_1(\omega)}\right)}{\Delta\phi} \tag{9}$$

式中：Q 为物质因子；$\Delta\phi=\dfrac{(n(\omega)-1)(d_1-d_2)\omega}{c}$ 为 2 个厚度对应的信号的相位差。可以看出，物质因子可以直接由 2 个不同厚度对应的信号的幅度和相位差直接求出而不需要参考信号，药品厚度及外包装等因素都会消去，因此可以用来对药品进行分类。

3 多特征联合检测的药品分类识别方法

目前，机器学习在药品识别中已有初步应用，主要是基于单个或 2 个特征的分类识别，有些药品的某一特征的差别不是很明显，采用单个特征进行判断会导致误判。因此本文将多个特征参数联合起来作为输入，采用 BP 神经网络、LVQ、SVM 三种机器学习方法分别对药品进行分类识别，以提高药品分类识别能力。

3.1 BP 神经网络

BP 神经网络[13]作为一种非线性的映射方式，能够将输入的特征值映射到网络的输出分类结果，实现从输入空间到输出空间的非线性映射，在模式识别和分类中具有良好的性能。根据 Kolrnogorov 定理，一个 3 层 BP 神经网络能够实现对任意非线性函数进行逼近，因此根据实际需要，本文采用由输入层、隐层和输出层组成的 3 层 BP 网络，如图 3 所示。BP 神经网络包括正向传播和误差的逆向传播。正向传播时，输入样本从输入层传入，经各隐层处理后，传向输出层。若输出层的实际输出与期望的输出不符，则误差将以某种形式通过隐层向输入层逐层反传。利用输出后的误差来估计输出层的直接前导层的误差，再用这个误差估计前一层的误差，如此一层一层地反传下去，就获得了所有其他各层的误差估计，此误差信号即作为修正各单元权值的依据。最后计算全局误差：

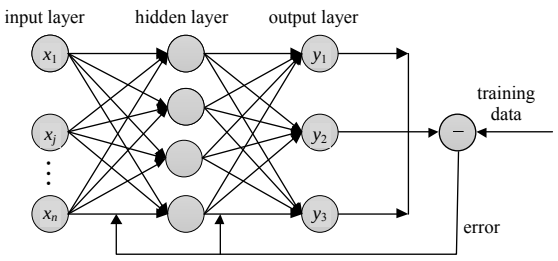

Fig.3 Structure of BP neural net

图 3 BP 神经网络

$$E=\frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{p}(d_o(k)-y_o(k))^2 \tag{10}$$

式中： m 为样本个数； p 为输出层神经元个数； d_o 为期望输出； y_o 为网络输出。

判断全局误差是否满足要求，当误差达到预设精确度或学习次数大于设定的最大次数，则结束算法；否则，选取下一个学习样本及对应的期望输出，进入下一轮学习，直到满足停止条件。

3.2 支持向量机

支持向量机(SVM)是一种基于结构风险最小化原则的分类模型，常用于模式识别、分类及回归分析[15−16]，在解决小样本、非线性及高维模式识别中表现出许多特有的优势。SVM 的基本思想是利用一个超平面将不同类别的样本分开，并且使样本集中所有数据到这个超平面的最短距离最大。超平面的方程可以写成：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0 \tag{11}$$

式中： $\boldsymbol{x}$ 为特征变量； $\boldsymbol{w}$ 为权矢量； b 为偏置。假设 $P=(x_1,x_2,\cdots,x_n)$ 为一个样本，其中 x_i 为第 i 个特征变量，该样本到超平面的距离 d 为：

$$d=\frac{\left|w_1^*x_1+w_2^*x_2+\cdots+w_n^*x_n+b\right|}{\sqrt{w_1^2+w_2^2+\cdots+w_n^2}}=\frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right|}{\|\boldsymbol{w}\|} \tag{12}$$

式中 $\|\boldsymbol{w}\|$ 为超平面的范数，因此目标函数可转化为：

$$\arg\max\left[\min\left(\frac{1}{\|\boldsymbol{w}\|}y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)\right)\right] \tag{13}$$

式中 y 为样本标签，当 $(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)\geqslant 1$ 时， $y=1$； $(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)\leqslant 1$ 时， $y=-1$，故问题可简化为： $\arg\max\left(\frac{1}{\|\boldsymbol{w}\|}\right)$。

将目标函数等价替换为：

$$\begin{cases}\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2\right)\\ \text{s.t. }(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)-1\geqslant 0\end{cases} \tag{14}$$

通过最小化目标函数可以求解超平面。对于非线性问题，通过引入核函数，可将其映射到高维空间进行求解。

3.3 学习向量量化

学习向量量化(LVQ)[12]是一种基于自组织神经网络的算法，是一种结构简单、功能强大的有监督式神经网络分类方法。LVQ 在训练过程中通过对神经元权向量不断更新，对其学习率的不断调整，能够使不同类别权向量之间的边界逐步收敛至贝叶斯分类边界。本文采用 LVQ1 算法，主要步骤包括：

1) 初始化输入层与竞争层之间的权值及学习率；
2) 将输入向量送达输入层，计算竞争层神经元与输入向量的距离；
3) 选择与输入向量距离最小的竞争层神经元；
4) 根据标签调整权值；
5) 循环训练，直到达到训练精确度或训练次数。

利用机器学习进行分类识别的过程如图 4 所示。首先利用上述方法对录取数据进行处理，分别提取折射率、吸收率等多个特征数据，然后将提取的特征数据排成一列作为训练样本。通过多次录取数据得到多个样本，并分成训练样本和测试样本。然后分别利用训练样本对 BP,SVM,LVQ 进行训练。最后利用测试样本对训练后的网络进行测试以评估分类识别效果。

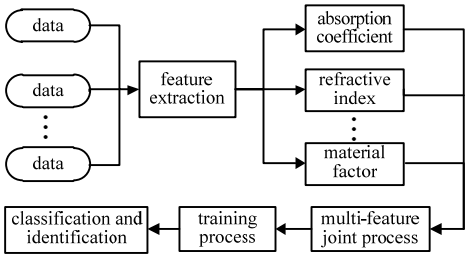

Fig.4 Flowchart of multi-feature joint detection method
图 4 多特征联合分类识别

4 实验结果及分析

4.1 实验数据与特征提取

测试的药品共有 3 种：克拉霉素、仁和可里克颗粒和板蓝根颗粒，分别用 sample1,sample2 和 sample3 表示。为保持药品的一致性，排除形状的影响，将 3 种药品取出干燥处理，捣碎成粉末状，装入同样的胶囊外壳中。实

验中以无药品时的太赫兹信号当做参考波形，放置药品时透过的太赫兹信号作为样品信号，时域信号和频域信号分别如图 5~6 所示。

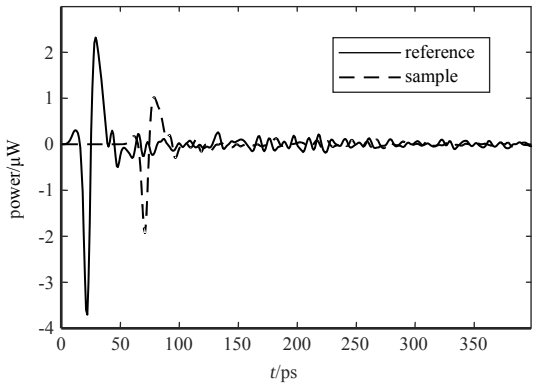

Fig.5 Time domain of reference and sample signals
图 5 参考和样本的时域信号

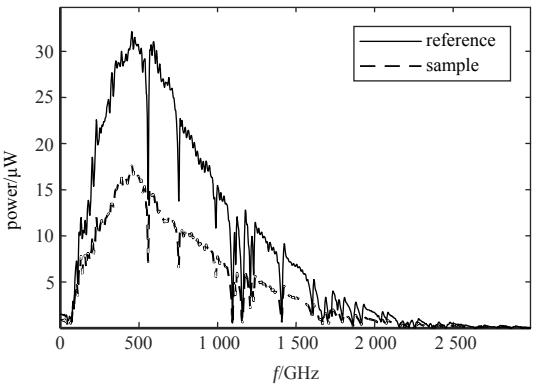

Fig.6 Spectrum of reference and sample signals
图 6 参考和样本的频域信号

药品的太赫兹光谱在低频段受系统条件的变化如激光功率的变化波动、校正和扫描数据时的电子噪声等的影响严重；在高频段的数据由于系统动态范围的限制会受到噪声的影响。因此，选择 0.2~1.4 THz 波段的数据作为数据的太赫兹响应对其进行分类检测。药品的其中一组折射率、吸收率分别如图 7~8 所示。由折射率可以看出，可立克颗粒和板蓝根颗粒比较近似，在 1.4~1.5 之间，而克拉霉素的折射率在 1.3~1.4 之间，与前面 2 种药品区别较明显。3 种药品的吸收率则可以区分。由此可见，不同特征的分辨能力不同，多特征联合能够提高分类识别效果。

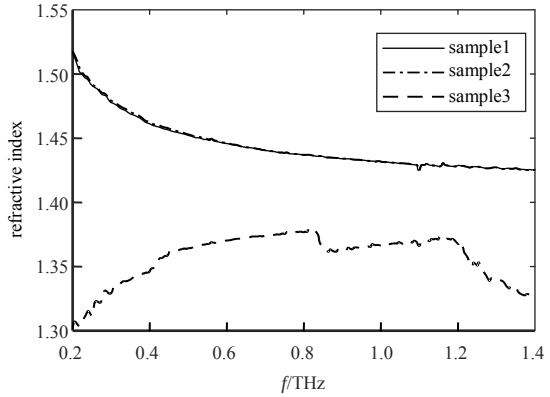

Fig.7 Refractive index of the three kinds of medicine
图 7 三种药品折射率

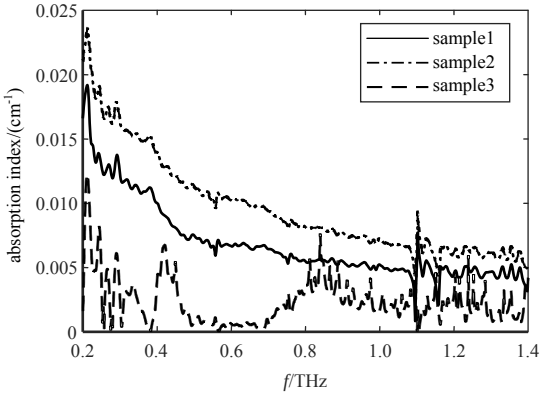

Fig.8 Absorption index of the three medicine
图 8 三种药品吸收率

4.2 基于多特征联合的机器学习分类识别结果

利用本文所述方法提取多个特征并组合成一列，可以得到一个样本。对每种药品不同位置测量 100 次，处理后得到 300 个样本。在小样本集中，一般选择 2/3~4/5 作为训练样本，其余为测试样本。实验随机选取 200 次作为训练样本，100 次作为测试样本。机器学习方法的学习效果容易受参数的影响，因此需要多次调节参数。在 BP 训练过程中，学习率设为 0.02，学习目标为 0.05，每次迭代循环次数为 100；在 LVQ 训练过程中，学习率设为 0.01，学习目标为 0.03，每次迭代循环次数为 100；在 SVM 训练过程中，采用 Matlab 自带的 svmtrain 函数进行训练。

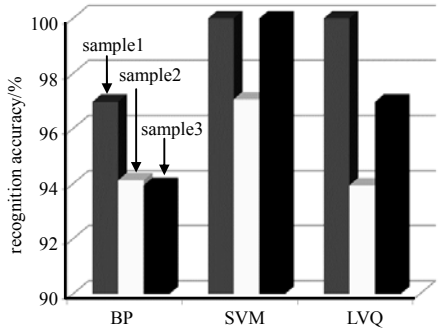

Fig.9 Comparison of the three medicine detection results
图 9 三种机器学习方法多特征检测结果对比

机器学习方法的学习效果还受训练样本的影响，因此采用随机选择训练样本并进行多次实验求平均，结果如表 1~3 所示。BP 神经网络识别率相对较低：克拉霉素、可立克、板蓝根识别准确率分别为 96.97%, 94.12%, 93.94%，平均识别准确率达到 95%；LVQ 准确率次之：克拉霉素、可立克、板蓝根识别准确率分别为 100%, 93.94%,

96.97%, 平均识别准确率达到 97%; SVM 准确率最高: 克拉霉素、可立克、板蓝根识别准确率分别为 100%, 97%, 100%, 平均识别准确率达到 99%。从分类识别结果可以看出, 克拉霉素的识别准确率最高, 说明克拉霉素与另外 2 种药品区别较明显。实验结果说明, 多特征联合具有很强的特征表征能力, 能有效准确地分类识别药品。3 种机器学习方法识别结果如图 9 所示, 可以看出 SVM 的分类识别准确率最高, 适用于进行药品检测。

表 1 BP 神经网络的分类识别结果

Table1 Performance of BP network

medicine	number	classification result			error	accuracy/%	average accuracy/%
		sample1	sample2	sample3			
sample1	33	32	0	1	1	96.97	
sample2	34	1	32	2	1	94.12	95
sample3	33	0	2	33	1	93.94	

表 2 SVM 方法的分类识别结果

Table2 Performance of SVM

medicine	number	classification result			error	accuracy/%	average accuracy/%
		sample1	sample2	sample3			
sample1	32	32	0	0	0	100	
sample2	34	0	33	1	1	97.06	99
sample3	34	0	0	34	0	100	

表 3 LVQ 神经网络的分类识别结果

Table3 Performance of LVQ

medicine	number	classification result			error	accuracy/%	average accuracy/%
		sample1	sample2	sample3			
sample1	34	33	0		1	100	
sample2	33	0	31	2	1	93.94	97
sample3	33	0	1	33	1	96.97	

5　结论

本文以 3 种药品为研究对象, 利用 THz-TDS 系统提取物品的谱吸收峰、折射率、吸收率、物质因子等多个特征参数, 然后分别采用 BP 神经网络、SVM 和 LVQ 3 种方法实现了多特征联合药品分类识别。实验中对基于多特征联合检测分类识别方法的 3 种机器学习方法的识别结果进行了比较,结果表明多特征联合模型对药品样品的平均识别准确率达到 95%以上, 其中 SVM 更是达到了 99%, 证明多特征联合具有很强的特征表征能力, 能够有效分类识别药品。本次实验结果表明, 基于太赫兹光谱的多特征联合模型能够有效解决药品分类识别问题, 对药品检测和分类识别具有重要的研究价值。

参考文献:

[1] 郑新,刘超. 太赫兹技术的发展及在雷达和通讯系统中的应用(Ⅱ)[J]. 微波学报, 2010,27(6):1-5. (ZHENG Xin,LIU Chao. Recent development of THz technology and its application in radar and communication system(Ⅱ)[J]. Journal of Microwaves, 2011,27(6):1-5.)

[2] 叶麾,郄明蓉,曹寒雨,等. 太赫兹技术在医学科学中的应用及研究进展[J]. 光电工程, 2018,45(5):170528. (YE Hui, QIE Mingrong,CAO Hanyu,et al. Application of terahertz technology in medical science and research progress[J]. Opto-Electronic Engineering, 2018,45(5):170528.)

[3] 任姣姣,李丽娟,张丹丹,等. 太赫兹无损检测的多特征参数神经网络分析技术[J]. 光子学报, 2017,46(4):204-210. (REN Jiaojiao,LI Lijuan,ZHANG Dandan,et al. Multi-feature parameter neural network analysis technique based on terahertz nondestructive testing[J]. Acta Photonica Sinica, 2017,46(4):204-210.)

[4] 赵中原. 基于 THz 技术的小麦品质无损检测研究[D]. 郑州:河南工业大学, 2016. (ZHAO Zhongyuan. A study on nondestructive testing for wheat quality based on THz technology[D]. Zhengzhou,China:Henan University of Technology, 2016.)

[5] 杨蕊,陈刚,蒋玲,等. 太赫兹光谱技术在纤维材料领域应用进展[J]. 应用激光, 2018,38(5):164-172. (YANG Rui, CHEN Gang, JIANG Ling,et al. Progress in the application of terahertz spectroscopy in the field of fiber materials[J]. Applied Laser, 2018,38(5):164-172.)

(下转第 254 页)