基于DCGAN的拉曼光谱样本扩充及应用研究

李灵巧[1,2]，李彦晖[2]，殷琳琳[4]，杨辉华[1,2*]，冯艳春[3]，尹利辉[3]，胡昌勤[3]

1. 北京邮电大学人工智能学院，北京 100876
2. 桂林电子科技大学计算机与信息安全学院，广西 桂林 541004
3. 中国食品药品检定研究院，北京 100050
4. 北京师范大学环境学院，北京 100875

摘 要 拉曼光谱检测方法依赖于化学计量学算法，深度学习是当下最炙手可热的方向，可应用于拉曼光谱进行建模。但是深度学习需要大样本进行训练，而拉曼光谱采集受制于器材和人力成本，获取大批量的样本需要更大成本，且易受荧光等因素干扰，这些问题都制约了将深度学习应用于拉曼光谱。针对以上问题，通过引入深度卷积生成对抗网络(DCGAN)提取拉曼光谱内部特征，对抗生成新的拉曼光谱，从而达到扩充数据集目的。同时和另一个扩充数据集的方法——偏移法进行对比，证明DCGAN的可靠性。设计生成光谱选取标准，选取高相似性的光谱填充数据集，为深度学习在拉曼光谱中的应用奠定基础。为了验证生成的光谱比原始光谱有更好的适用性，设计四组实验：(1)使用原始拉曼光谱输入到SVM进行分类，得到51.92%的分类准确率；(2)使用原始拉曼光谱输入到CNN进行分类，得到75.00%的分类准确率；(3)采用偏移法生成光谱，输入到CNN里进行分类，得到91.85%的分类准确率；(4)使用DCGAN生成光谱，输入到CNN里进行分类，得到98.52%分类准确率。实验结果表明，DCGAN能在只有少量拉曼光谱的情况下，通过对抗学习得到较好的生成光谱，且生成的光谱相比原光谱更加清晰，减少了可能的干扰因素，具有光谱预处理效果。通过DCGAN对抗生成大量高质量的数据填充到原有拉曼光谱数据集，扩充数据集的样本量，使得深度学习模型能够得到更好的训练，从而提高模型的准确率。该研究为深度学习方法应用于拉曼光谱分析技术提出了一个可行的方案。

关键词 拉曼光谱；数据扩充；光谱分类；深度卷积生成对抗网络

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)02-0400-08

引 言

食品药品的安全一直是人们重点关注的对象，常用的食品药品检测手段有吸收系数法、化学法和HPLC法等，不仅繁琐，而且局限于实验室。因此需要一种可以快速检测的手段，近年来发展较好的是近红外光谱检测和拉曼光谱检测。拉曼检测技术是基于拉曼光谱特征位移峰而产生的一种检测技术。当光照射到物体分子上时会发生弹性散射，额外会有少量光子发生非弹性散射，这些光子就是拉曼光子，拉曼光子转移能量到分子上，产生位移散射光，位移的距离对应分子的信息。不同的距离长短对应了不同的分子结构，由此产生拉曼谱图。根据谱图就可以明确样品化学与分子信息和含量[1]。相比红外光谱法，拉曼光谱提供的是无损定性定量分析，对样品无特殊要求，短时简便高灵敏度，避免了因为样品的破坏或者样品自身的缺陷导致的误差[2]。

由于仪器和方法的改进升级，使用拉曼分析对食品药品进行鉴别和分类得到了广泛应用。目前主流的分类算法线性学习机(linear learning machine，LLM)、软独立建模分类法[3](soft independent modeling of class analogy，SIMCA)、人工神经网络[4](artificial neural network，ANN)、K-最近邻[5](K-nearest neighbor method，KNN)等。最近两年，我们将浅层机器学习方法应用于近红外光谱药品分类[6]，并取得了较好的分类结果。这些方法各有优点但较为传统，目前深

收稿日期：2020-02-05，**修订日期**：2020-06-02

基金项目：国家自然科学基金项目(61906050)，广西科技计划项目(2018AD11018)，桂林电子科技大学研究生教育创新计划项目(2018YJCX44)资助

作者简介：李彦晖，1994年生，桂林电子科技大学计算机与信息安全学院硕士研究生 e-mail：1703201023@mails.guet.edu.cn

*通讯作者 e-mail：yhh@bupt.edu.cn

度学习方法在图像分割[7]、图像增强[8]和图像检测[9]等方面大放异彩，将深度学习应用到光谱学是必然趋势。现有拉曼光谱采集需要较高的人力和时间成本，采集到的数据样本量较少和存在干扰因素，不能满足深度学习需要用大样本进行训练的条件，因此将深度学习应用在拉曼光谱中的研究较少。

鉴于此，本文提出一种将深度学习应用到拉曼光谱的方法：使用深度卷积生成对抗网络[10](deep convolutional generative adversarial networks，DCGAN)生成新光谱，并输入CNN进行分类。目前GAN在光谱分析中应用不多，仅见应用于高光谱分析，而在拉曼和近红外光谱分析方面未见报道。

在搭建深度学习模型的过程中，常遇到因训练数据集样本量不够导致欠拟合的问题。解决该问题除了在算法层面的优化，还需拓展训练集样本数量。常用的数据增强方法有形状变换、监督式抠取、GAN等。本文采取的DCGAN则是在原始GAN的基础上引入卷积，借助卷积层的特征提取能力，提取拉曼光谱的深层特征，生成高度相似的光谱。

采用DCGAN扩充拉曼光谱，扩充训练集样本量并提升CNN分类精度。设置数据增强扩充光谱并输入CNN进行分类，与DCGAN的结果进行对比。实验结果表明：DCGAN生成的光谱能够被CNN识别并进行分类，增加的数据集样本量提升了CNN的分类精度。其次，DCGAN可以实现使用少量原始拉曼光谱对抗生成新光谱，达到扩充数据集的样本量目的，有效减少人力和时间成本。

1　算法描述

1.1　CNN(卷积神经网络)

1.1.1　算法介绍

CNN通常包含卷积层、池化层、全连接层，先正向传播得到输入数据特征，然后反向传播使用梯度下降进行迭代，完成权值更新。

卷积层通过卷积运算提取输入数据特征，卷积公式如式(1)

$$x_\beta^\gamma = f\left(\sum_{\alpha\in M_\beta} x_\alpha^{\gamma-1} k_{\alpha\beta}^\gamma + b_\beta^\gamma\right) \tag{1}$$

式(1)中：$x_\alpha^{\gamma-1}$ 和 x_β^γ 分别是第 $\gamma-1$ 层、第 γ 层输出特征上相应的值；$k_{\alpha\beta}^\gamma$ 为卷积核的权重值；b_β^γ 为特征的偏置；f 是卷积层神经元的激活函数。

池化层对卷积层提取得到的特征进行进一步降维，加快运算速率。池化层的公式如式(2)

$$x_\beta^\gamma = \text{pool}(x_\beta^{\gamma-1}) \tag{2}$$

式(2)中：$x_\beta^{\gamma-1}$ 和 x_β^γ 分别是第 $\gamma-1$ 层、第 γ 层输出特征上相应的值；pool表示在最大池化和平均池化里选择的函数。

1.1.2　改进的CNN

卷积神经网络主要用于图像分类，输入一般为 $n\times n$ 维的图像，对应卷积核及池化操作均是 $n\times n$ 的矩阵，并不适用于光谱，需要针对光谱对网络进行修改。这里修改卷积核尺寸为 1×5。光谱谱线中最重要的是每个波长点的峰强信息，然而CNN里的池化操作会使得光谱信息大量丢失并不利于分析，所以这里舍弃池化层。同时为了减小运算量，将网络输出层和中间层修改为单层感知器。经过改进后设计为一个5层的CNN网络，具体网络结构如表1。

表1　CNN网络各层设计

Table 1　CNN network design

网络层	设置参数
INPUT	预处理后的拉曼光谱数据
Conv1	Size：1×5，stride：1，ReLU
Conv2	Size：1×5，stride：1，ReLU
Full Connected	Conv2的feature展开
OUTPUT	9个输出神经元，连接FC层

1.2　DCGAN(深度卷积生成式对抗网络)

1.2.1　算法介绍

DCGAN的网络结构如图1：图中左边是G(Generator)网络，右边是D(Discriminator)网络。

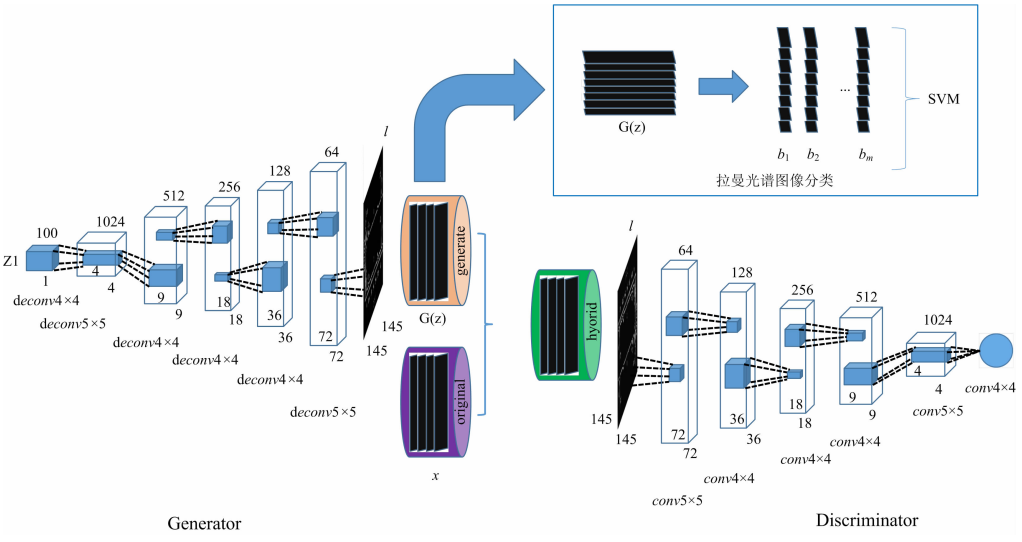

图1　用于拉曼光谱分类的DCGAN网络结构示意图

Fig. 1　Diagram of DCGAN network structure for Raman spectrum classification

G 是生成网络，给它输入一个随机噪声 z，最终能生成一张图片，标记为 $G(z)$。

D 是判别网络，用来判别某张图片真实的程度。给它输入一张图片 x，会输出 $D(x)$，代表 x 是真实图片的概率，若概率数值是 1 就说明图片完全真实。若概率数值为 0 就说明图片作假。

DCGAN 引入卷积计算图像整体区域特征信息，从而具有很强特征提取能力。由于卷积网络中池化层(pooling)的下采样会造成图像信息部分损失，不能采用，因此把 G 和 D 网络中的池化层替换为反卷积层和步进卷积层，减少图像信息损失。然后引入 Batch Normalization (BN)构造更加稳定的网络。

1.2.2 改进的卷积层

传统的 DCGAN 网络的卷积层主要面向图像分类为主。该网络层默认输入一般是二维图像，因此网络层的卷积核和池化窗口都是大小为 $n\times n$ 维的矩阵。如此来看这样的网络结构并不适用于光谱数据，因此需要对传统 DCGAN 网络的卷积层进行改进，也就是将 DCGAN 中卷积层的卷积核修改为一维向量卷积核，使之能够处理拉曼光谱数据。

1.2.3 DCGAN 网络结构设计

针对 Raman 光谱数据设计的生成网络和判别网络的结构见表 2。

表 2 用于 Raman 光谱扩充的生成网络和判别网络

Table 2 Generator network and Discriminator network for Raman spectral augmentation

生成网络						判别网络					
网络层	卷积核	stride	padding	激活函数	BN 层	网络层	卷积核	stride	padding	激活函数	BN 层
deconv1	1×4	0	0	ReLU	是	conv1	1×5	2	1	LeakyReLU	是
deconv2	1×5	2	1	ReLU	是	conv2	1×4	2	1	LeakyReLU	是
deconv3	1×4	2	1	ReLU	是	conv3	1×4	2	1	LeakyReLU	是
deconv4	1×4	2	1	ReLU	是	conv4	1×4	2	1	LeakyReLU	是
deconv5	1×4	2	1	ReLU	是	conv5	1×5	2	1	LeakyReLU	是

1.3 数据增强

仅使用 DCGAN 生成光谱来进行分类缺少算法效果对照。增加一个数据增强方法生成光谱，通过两种方法对生成的光谱进行分类对比。

数据增强是一个扩展数据集最常用的技术，它已成功地应用于许多领域，从图像分类到分子建模。其核心思想是通过模拟数据集中的各种数值变化，从有限的标记样本数目中扩展训练样本的数目。对于光谱数据，采用随机偏移量、斜率的随机变化和随机乘法来扩展数据集。偏移量为训练集标准差的±0.10 倍，叠加次数为训练集标准差的 1±0.10 倍，斜率在 0.95～1.05 之间均匀随机调整。其函数表达如式(3)和式(4)

$$x' = kx + b \tag{3}$$

$$b = ma + n - a - \frac{m}{2} + 0.5 \tag{4}$$

式中，k 为缩放比例，b 为偏移项，表示对光谱每个数据点随机向上偏移，每个点的偏移量呈线性递增或递减形式。m 为倾斜度，n 为倾斜时的偏移，a 为步长从 0 到 1 之间的向量。x 表示原光谱，x'表示用 x 生成的光谱。

图 2 是数据增强生成光谱的示意图，图中粗蓝线为原始光谱，其余为偏移法生成光谱。

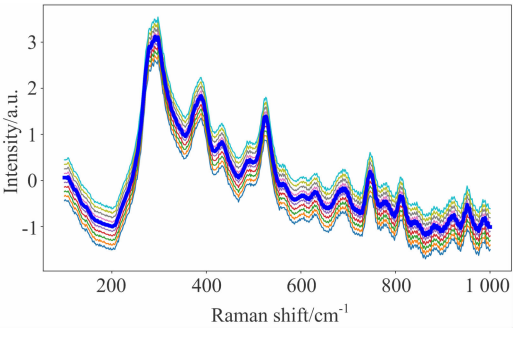

图 2 偏移法扩增光谱示意图

Fig. 2 Spectral augmentation by slope-bias adjusting

1.4 分类方法

分类方法有以下几种：无监督分类、半监督分类、有监督分类。常见的无监督分类算法有 K 聚类、Fuzzy Means[11]；半监督学习则是 DBSCAN 最常用；对于有监督分类来说，常用的有支持向量机(support vector machine，SVM)。

根据需要选择分类方法，CNN 上文已提到不再赘述，增加一个机器学习分类方法作为 CNN 分类方法参照，这里选用 SVM 方法。生成的拉曼光谱数据表示为 $G(z)=[b_1, b_2, \cdots, b_m]$，$m$ 为样本总数，SVM 的分类函数的对偶形式表示为

$$\max_{a}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(b_i, b_j)$$

$$\text{s.t.}\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{5}$$

$$0 \leqslant \alpha_i \leqslant C,\ i = 1,2,\cdots,m$$

其中 $k(b_i, b_j)$本文选择径向基函数(RBF)

$$k(b_i, b_j) = \exp\left(-\frac{\| b_i - b_j \|^2}{2\sigma^2}\right) \tag{6}$$

建模选择 LibSVM 软件，这里有两个参数 c 和 g，c 就是式(5)中的 C，$g=1/2\sigma$，参数设置为：$c=200$，$g=0.01$。

1.5 模型评价方法

1.5.1 扩充样本的选取标准

对生成图像进行评估有一定的困难，一般只能通过人工样本筛选和主观判断的方法来进行评价，不仅耗时而且费力。结构相似度(structural similarity index，SSIM)指标能够

很好的判断两个样本的相似性，故引入该指标对生成光谱进行评判。见式(7)[12]

$$\mathrm{SSIM}(x,g)=\frac{(2\mu_x\mu_g+c_1)(2\sigma_{xg}+c_2)}{(\mu_x^2+\mu_g^2+c_1)(\sigma_x^2+\sigma_g^2+c_2)} \tag{7}$$

式中 μ_x，μ_g，σ_x，σ_g 为 x 和 g 的均值和方差，σ_{xg} 为 x 和 g 协方差。c_1 和 c_2 为常数，用来保证函数稳定性，$c_1=(k_1L)^2$，$c_2=(k_2L)^2$，$L=255$，是图像像素最大值，$k_1=0.01$，$k_2=0.03$。

SSIM 取值范围[0，1]，大小与图像相似度成正比。这里设置 SSIM 阈值为 0.9，因为高相似度的生成光谱才能用于样本扩充。计算原始光谱和生成光谱之间 SSIM 值，SSIM 值大于等于 0.9 才采用该生成光谱，否则不用。

1.5.2　模型分类结果的评价方法

本实验采用分类准确率定量评价分类结果。

当光谱输入到分类器时，计算其分类的准确率 P。分类准确率 P。可表示为

$$P=\frac{N_c}{N_r}\times 100\% \tag{8}$$

式中：N_c 为正确分类的样本数，N_r 为样本数。

2　实验部分

2.1　数据处理

实验中使用的数据为中国食品药品检定研究院测取的药品拉曼光谱数据集，选取 9 类药品，药品分布如表 3。测量仪器为同方威视 RT6000、Metage OPAL 3000 和 Opto Trace RamTracer-200-HS 拉曼光谱仪，测量参数 Metage OPAL 3000 和 Opto Trace RamTracer-200-HS 积分时间设为 25 s，积分次数设为 3 次，同方威视 RT6000 积分时间设为 25 s，积分次数为自动。为了避免实验中因为样本波段不一致而导致的结果不理想，以下实验均选择每种药品在 100～1 000 cm^{-1} 的光谱。同时为了验证 DCGAN 在生成光谱中具有预处理的作用，实验所采用的所有光谱均只进行基线校正和归一化预处理，为减少计算量，采样间隔选择隔 13 点采样。

表 3　中检院数据集对应的药品分布

Table 3　Distribution of corresponding drugs in data set National Institute for Food and Drug Control

序号	药品类别	样本数目
1	Gatifloxacin	15
2	Lomefloxacin	21
3	Norfloxacin	15
4	Pefloxacin	12
5	Cephradine	18
6	Cefradline	15
7	Cefixime	9
8	Ceftazidime	18
9	Cefdinir	30

2.2　CNN 模型训练

卷积神经网络学习率设置为 0.001，梯度更新块大小设置为 32。训练过程中手动调整以保持所有层具有相同的迭代速度。对卷积层设置权值初始化为 0.01 标准差的零均值高斯分布。对全连接层的权重设置 0.005 标准偏差。由于卷积后的结果会导致光谱首尾数据的丢失，因此输入前对原始光谱采用 0 填充。目标函数采用最小化预测值和真值的交叉熵

$$\min\left\{-\frac{1}{N}\sum_{i=1}^{N}[y_i\ln\hat{y}_i+(1-y_i)\ln(1-\hat{y}_i)]\right\} \tag{9}$$

式(9)中，N 为样本数，y_i 是样本 i 的类别标记，$\hat{y}_i$ 是其预测结果。

2.3　DCGAN 模型训练

DCGAN 中卷积网络的激活函数选择 LeakyReLU，设定 leak 的斜度值为 0.2，整个网络设定数值为 2 的 batch size，网络的学习率不能太大，否则时间过长，这里设置为 0.000 2，卷积层还需要使用优化器并设置动量参数，优化器使用 Adam，参数设置为 0.5 时可以稳定训练。D 训练两次，G 训练一次，迭代次数设置为 800。每迭代 10 次输出一次 SSIM 的平均值，作为选取扩充样本的标准。

将原始拉曼光谱作为初始数据集，通过对抗生成新的数据集，为了有所区别，这里给生成网络设定 100 个服从标准正态分布的噪声 z，通过反卷积网络后能够生成和真实图像相似的“假”样本。然后将真假样本同时输入判别网络，通过卷积层能够得到范围为 0 到 1 的概率值，根据概率值判断样本的真假程度。训练分为两个部分：

(1)训练生成网络，提前设定好判别参数，用以优化生成网络，直到生成的“假”样本判别网络无法识别，此时生成网络输出大概率真实的样本，映射到函数内就是最大化 $D(G(z))$，亦即最小化 $1-D(G(z))$。

(2)训练判别网络，类似地，给定生成网络的参数，区性优化判别网络，这样能大大提高判别网络的精度，这里期望最大化 $D(x)$。生成样本 $G(z)$需要使得 $D(G(z))$最小。对判别网络的目标函数优化为 $\ln D(x)+\ln(1-D(G(x)))$。

最终得到目标函数

$$\min_G\max_D V(D,G)=E_{x\sim P_{\mathrm{data}(x)}}[\ln D(x)]+E_{z\sim P_z(z)}[\ln(1-D(G(z)))] \tag{10}$$

然后固定生成网络的参数，以优化判别网络，使得 $V(D,G)$ 最大

$$\max V(D,G)=\int_x[P_{\mathrm{data}}(x)\ln(D(x))+P_g(x)\ln(1-D(x))]\mathrm{d}x \tag{11}$$

为了式(11)最大，这需要式(12)

$$P_{\mathrm{data}}(x)\ln(D(x))+P_g(x)\ln(1-D(x)) \tag{12}$$

取得最大值。显然有：对任意非零的 $P_{\mathrm{data}}(x)$，$P_g(x)$，且实数值 $D(x)\in[0,1]$时，式(12)在 $P_{\mathrm{data}}(x)/(P_{\mathrm{data}}(x)+P_g(x))$ 处取得最大值，列出最优的生成网络 D 的函数

$$D_G(x)=\frac{P_{\mathrm{data}}(x)}{P_{\mathrm{data}}(x)+P_g(x)} \tag{13}$$

对生成网络进行优化时，有 $P_{\mathrm{data}}=P_g$ 时生成网络取得最优解，使得生成网络更好地再现真实样本的分布。

3 结果与讨论

设计四组实验。分别是 SVM 对原始光谱进行分类的参照组、CNN 对原始光谱进行分类的对照组、CNN 对 DCGAN 生成光谱进行分类的实验组和 CNN 对偏移法生成光谱进行分类的实验组。

由于 SVM 和 CNN 需要进行训练，在原始光谱的实验中，选取 70%作为训练集，剩下 30%作为测试集。在生成光谱的实验中，分别用数据增强方式和 DCGAN 方式将每种药品的谱图数扩充到 100 个，再选取 70%的光谱对 CNN 进行训练，剩下 30%的光谱进行测试。

3.1 原始光谱直接 SVM 分类

表 4 展示了对原始光谱的训练集和测试集的划分情况，SVM 分类原始拉曼光谱的结果如表 5 所示。以 SVM 直接对原始的拉曼光谱分类产生的结果来看，对拉曼光谱的分类准确率并不高。对于 Pefloxacin 和 Cefixime 两个样本最少的药品来说，几乎无法准确分类。从表 5 中可以看出由于总体数据集样本量不大，其分类精度依然有待提高。

表 4 药品样本训练集、测试集划分情况

Table 4 The training set, test set distribution of drug samples

药品	训练集	测试集
Gatifloxacin	11	4
Lomefloxacin	15	6
Norfloxacin	11	4
Pefloxacin	9	3
Cephradine	13	5
Cefradline	11	4
Cefixime	7	2
Ceftazidime	13	5
Cefdinir	21	9
总数	111	52

表 5 中检院拉曼光谱数据判别详细结果-SVM(%)

Table 5 Detailed results of Raman spectrum discrimination of China food and drug institute-SVM (%)

药品	分错数量	
	训练集	测试集
Gatifloxacin	5	3
Lomefloxacin	4	2
Norfloxacin	4	2
Pefloxacin	7	3
Cephradine	4	3
Cefradline	5	3
Cefixime	6	2
Ceftazidime	4	4
Cefdinir	2	3
分类准确率	63.06(70/111)	51.92(27/52)

3.2 原始光谱直接 CNN 分类

用于 CNN 分类的训练集和测试集的划分同表 4。表 6 展示了相同波段的原始光谱输入 CNN 的分类实验结果。表中分别列出了 CNN 对原始拉曼光谱分类结果(训练集分类准确率 78.38%，测试集分类准确率 75.00%)。同样对于 Pefloxacin 和 Cefixime 两个样本最少的药品来说，分类准确率略有提升。由于 CNN 具有很强的特征提取和分类能力，因此 CNN 对拉曼光谱的总体分类精度高于 SVM。

表 6 拉曼光谱数据判别详细结果-CNN(%)

Table 6 Detailed results of Raman spectrum discrimination CNN (%)

药品	分错数量	
	训练集	测试集
Gatifloxacin	3	2
Lomefloxacin	2	1
Norfloxacin	2	2
Pefloxacin	4	2
Cephradine	2	1
Cefradline	3	2
Cefixime	5	2
Ceftazidime	2	1
Cefdinir	1	0
分类准确率	78.38(87/111)	75.00(39/52)

3.3 数据增强生成光谱直接 CNN 分类

上述实验仅用原始光谱进行分类对比实验，为了实验的严谨性，需要考虑到生成的光谱是否具有相同的优越性。因此需要扩增谱图和原始谱图分类对比来查看情况，先用偏移法将每个药品光谱扩充到 100 个，训练集和测试集划分见表 7。分别输入 CNN 训练并分类得到结果，图 3 为单个药品生成 10 个谱图和原始谱图的叠加图。表 8 是生成光谱数据判别详细情况。实验结果表明，偏移法生成谱图具有较好的分类准确率；另一方面，分类过程中出现了一些误分类的情况，即把本该分类到某种药品的谱图认为是另一种药品的谱图。出现这种现象的原因是偏移法生成的光谱有些波长点的峰强信息会改变，此时该拉曼峰可能会被认为是另一种分子。同时，由于偏移法生成光谱是对原光谱的重塑，因此有必要评估生成光谱相比原光谱的失真度。局部方差估计法

表 7 药品样本训练集、测试集划分情况

Table 7 The training set, test set distribution of drug samples

药品	训练集	测试集
Gatifloxacin	70	30
Lomefloxacin	70	30
Norfloxacin	70	30
Pefloxacin	70	30
Cephradine	70	30
Cefradline	70	30
Cefixime	70	30
Ceftazidime	70	30
Cefdinir	70	30
总数	630	270

LVE(local variance estimation method)是一个较好的能够估计图像失真程度的方法，其算法原理是先计算每张图片像素局部方差，最大的局部方差为信号方差，最小的局部方差为噪声方差，计算信号方差和噪声方差的比值，并转换成dB表示。表 9 为图 3 生成的 10 个谱图对比原始谱图的信噪比，从实验结果可以看出生成光谱相对原始光谱的失真程度。

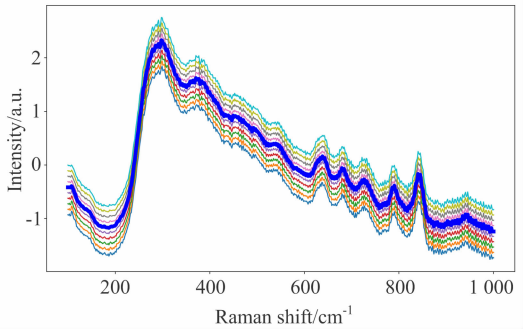

图 3　偏移法生成谱图

Fig. 3　Spectral generation by data augmentation

表 8　中检院拉曼光谱数据判别详细结果-偏移法＋CNN(%)

Table 8　Detailed results of Raman spectrum discrimination of China food and drug institute-Data augmentation and CNN (%)

药品	分错数量	
	训练集	测试集
Gatifloxacin	9	1
Lomefloxacin	5	3
Norfloxacin	8	2
Pefloxacin	10	2
Cephradine	4	2
Cefradline	8	4
Cefixime	7	1
Ceftazidime	8	3
Cefdinir	7	4
分类准确率	89.52(564/630)	91.85(248/270)

表 9　偏移法扩增谱图的 LVE 方法信噪比(对应图 3)

Table 9　LVE signal to noise ratio of augmented spectral by slope-bias adjusting (corresponding to Fig. 3)

原始谱图(a)	扩增谱图(b)
24.56	31.24
	29.74
	30.18
	32.05
	31.77
	34.65
	28.47
	30.69
	29.69
	31.51

3.4　DCGAN 生成光谱直接 CNN 分类

同样使用 DCGAN 将每个药品光谱数量扩充到 100，训练集与测试集划分同表 7。图 4 为原始的中检院数据中随机选取的 10 个药品光谱如图 4(a)，和 DCGAN 进行对抗生成的 10 个新的光谱如图 4(b)的展示。从视觉上看出生成的光谱相较原始光谱更加平滑清晰，说明 DCGAN 在生成光谱的过程中能够起到预处理的作用。将划分好的训练集和测试集输入 CNN 进行训练分类，得到如表 10 所示的判别结果。实验结果表明生成谱图具有高分类准确率。同样评估 DCGAN 生成光谱相比原光谱的失真度，实验结果见表 11，从实验结果可以看出生成光谱的失真程度对比原始光谱差异较小，相比偏移法，DCGAN 生成的光谱较好的保留了原始谱图的信息。图 5 是四个实验的训练集和测试集的分类准确率对比，从分类准确率来看，DCGAN 生成的拉曼光谱数据更有利于准确分类。

表 10　拉曼光谱数据判别详细结果-DCGAN＋CNN(%)

Table 10　Detailed results of Raman spectrum discrimination of DCGAN and CNN (%)

药品	分错数量	
	训练集	测试集
Gatifloxacin	3	1
Lomefloxacin	4	0
Norfloxacin	2	2
Pefloxacin	3	0
Cephradine	1	0
Cefradline	3	0
Cefixime	2	0
Ceftazidime	3	0
Cefdinir	2	1
分类准确率	96.35	98.52

表 11　DCGAN 扩增谱图的 LVE 信噪比(对应图 4)

Table 11　LVE signal to noise ratio of augmented spectral by DCGAN (corresponding to Fig. 4)

原始谱图(a)	扩增谱图(b)
27.67	29.28
27.08	30.59
28.65	31.63
20.49	31.79
29.01	32.02
27.53	31.67
20.65	29.28
28.22	29.32
21.56	30.37
20.86	29.34

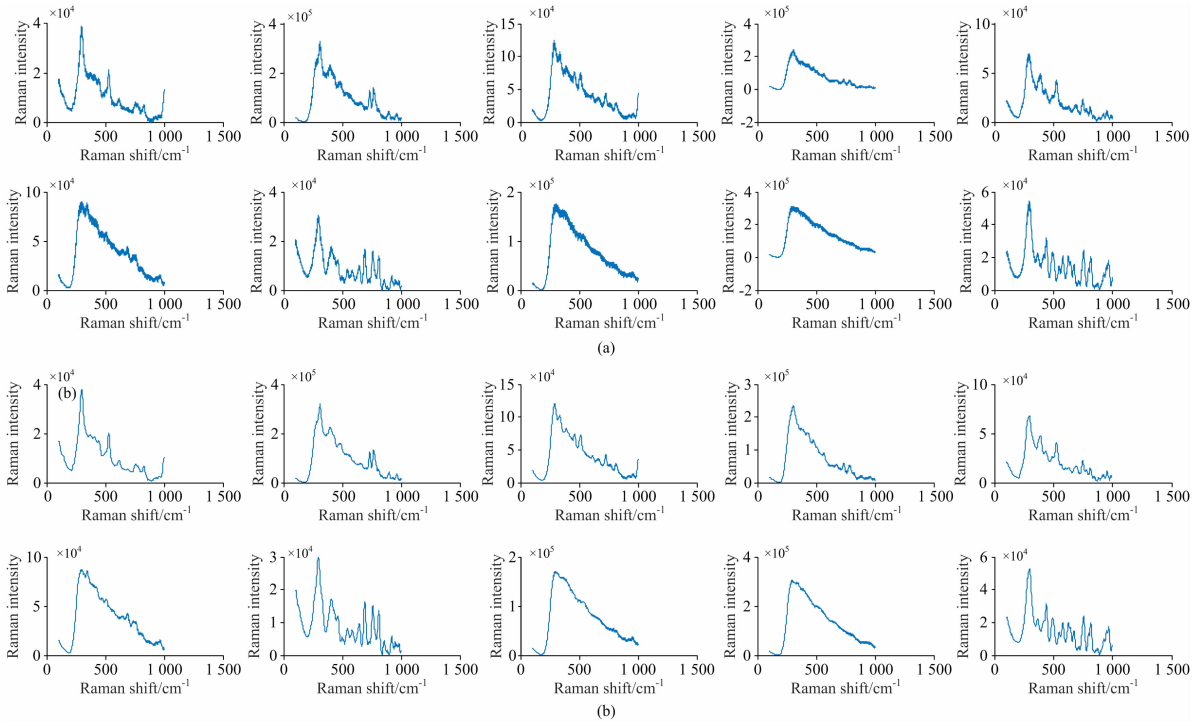

图 4 原始谱图(a)和 DCGAN 生成谱图(b)对比

Fig. 4 The original spectra (a) were compared with the generated spectra (b) of DCGAN

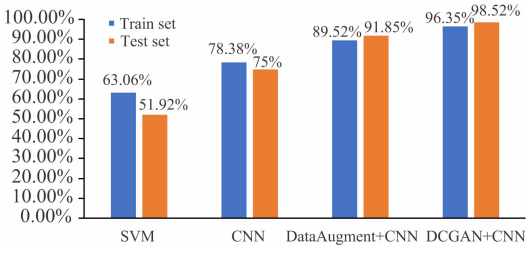

图 5 训练集和测试集的分类准确率对比图

Fig. 5 Comparison of classification accuracy of training set and test set

4 结 论

本文提出的基于 DCGAN 的数据扩充可有效扩充Raman光谱数据,并可由此提高对扩充后数据分类的准确率。使用中检院的药品拉曼光谱数据集进行实验,实验结果表明:

(1)由于中检院药品数据集样本量不大,该方法实现由少量的拉曼光谱生成更多的拉曼光谱扩充数据集,对解决由于数据集样本量不够而引发的深度学习分类精度较低的问题提出了一个新的思路;

(2)使用 DCGAN 网络对光谱进行生成甚至超分辨率重建是可行的,并且效果较好;

(3)基于 DCGAN 左右互搏的思想,使得生成的光谱图和原始光谱图在不断互相"欺骗"的过程中,提高了模型对光谱特征的识别和分类精度。

本文提出方法也存在不足之处,例如 DCGAN 是对原谱图进行重塑,不能应用常用的图像评价指标,因此如何更直观地反映生成谱图和原谱图的关系还有待研究;另一方面由于数据集的样本量不够大,在大样本容量时的实验结果仍需进一步验证。

References

[1] XU Lin-nan, LIN Hong, NIU Bing, et al(许林楠, 林 泓, 钮 冰, 等). Journal of Instrumental Analysis(分析测试学报), 2019, (11): 1400.

[2] LI Jia-jia, LIU Jing-li, JIN Ru-yi, et al(李佳佳, 刘靖丽, 靳如意, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(8): 2403.

[3] Lee Yonghoon, Han Song-Hee, Nam Sang-Ho. Applied Spectroscopy, 2017, 71(9): 2199.

[4] Sesmero M P, Alonso-Weber J M, Sanchis A. Information Fusion, 2020, 58: 132.

[5] Anzanello M J, Ortiz R S, Limberger R. Forensic Science International, 2014, 235(2): 1.

[6] ZHANG Wei-dong, LI Ling-qiao, HU Jin-quan, et al(张卫东, 李灵巧, 胡锦泉, 等). Chinese Journal of Analytical Chemistry(分析化

学)，2018，46(9)：1446.
[7] Pan X，Li L，Yang H，et al. Neurocomputing，2017，229：88.
[8] Zhang Weidong，Dong Lili，Pan Xipeng，et al. IEEE Access，2019，7(1)：72492.
[9] Pan Xipeng，Yang Dengxian，Li Lingqiao，et al. World Wide Web-Internet and Web Information Systems，2018，21(6)：1721.
[10] Radford A，Metz L，Chintala S. Computer Science，2015，47(8)：169.
[11] Polyakov A E，Ivanov M S. Fibre Chemistry，2018，49(6)：405.
[12] Theagarajan R，Bhanu B. PLOS One，2019，14(3)：e0212849.

Data Augmentation of Raman Spectral and Its Application Research Based on DCGAN

LI Ling-qiao[1,2]，LI Yan-hui[2]，YIN Lin-lin[4]，YANG Hui-hua[1,2*]，FENG Yan-chun[3]，YIN Li-hui[3]，HU Chang-qin[3]
1. School of Artificial Intelligence，Beijing University of Posts and Telecommunications，Beijing 100876，China
2. Man-Machine Intelligence Laboratory，Guilin University of Electronic Technology，Guilin 541004，China
3. National Institutes for Food and Drug Control，Beijing 100050，China
4. School of Environment，Beijing Normal University，Beijing 100875，China

Abstract The detection method of Raman spectroscopy relies on the chemometrics algorithms，and deep learning is the most popular are at present，which can be applied to the modeling of Raman spectroscopy. However，deep learning requires large samples for training，while Raman spectral collection is limited by equipment and labor cost. Obtaining large quantities of samples requires a higher cost，and also is suffered by fluorescence and other factors，which all restrict the application of deep learning to Raman spectral. In view of the above problems，the paper introduces the deep convolution generation counter network (DCGAN) to extract the characteristics of Raman peaks in the Raman spectrum，and generates a new Raman spectrum to expand the data set. At the same time，the reliability of DCGAN was proved by comparing with the slope-bias adjusting method，another method to expand the data set. In this paper，spectral selection criteria are designed and generated to fill the dataset with highly similar spectra，which is the first step for the application of deep learning in Raman spectra. In order to demonstrate that the generated spectrum has good comformality with the original spectrum，the paper sets up four groups of experiments for comparison：(1) the original Raman spectrum is input to SVM for classification，and the classification accuracy is 51.92%，(2) the original Raman spectrum was input to CNN for classification，and 75.00% classification accuracy was obtained，(3) the slope-bias adjusting method was used to generate the spectrum，which was input into CNN for classification，and the classification accuracy of 91.85% was obtained，(4) DCGAN was used to generate the spectrum，which was input into CNN for classification，and the classification accuracy was 98.52%. The comparison of the four groups of results proves the superiority of the Raman spectrum generated by DCGAN. The experimental results show that DCGAN can generated much alike spectrum through antagonism learning with only a small amount of Raman spectrum，and the generated spectrum is clearer than the original spectrum，reducing some interference factors，and has a preprocessing effect on the spectrum. Taking the advantage of DCGAN，a large number of high-quality data can be generated and filled into the original Raman spectral data set，and the sample size of the data set can be expanded，so that the deep learning model could be better trained，thus improving the accuracy of the classification or other model. This paper proposes a feasible scheme for applying deep learning method to Raman spectroscopy.

Keywords Raman spectrum；Data augmentation；Spectral classification；Deep convolutional generative adversarial networks

(Received Feb. 5，2020；accepted Jun. 2，2020)

* Corresponding author