深度卷积网络的多品种多厂商药品近红外光谱分类

李灵巧[1,2]，潘细朋[1]，冯艳春[3*]，尹利辉[3]，胡昌勤[3]，杨辉华[1,2*]

1. 北京邮电大学自动化学院，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

摘 要 近红外光谱(NIR)分析具有分析高效、样品无损、环境无污染以及可现场检测等优点，特别适合药品的快速建模分析。但NIR存在吸收强度弱以及谱带重叠等缺点，需要建立稳健可靠的化学计量学模型对其进行分析。深度卷积神经网络是深度学习方法中一个重要分支，它通过逐层抽取数据特征并进行组合、转换，形成更高层的语义特征，具有极强的建模能力，广泛应用于计算机视觉、语音识别等领域，而在药品NIR分析方面尚未见报道。基于深度卷积网络模型，对药品NIR多分类建模进行研究。针对药品NIR数据的特点，设计若干个面向多品种、多厂商药品NIR分类的一维深度卷积网络模型。模型中卷积层和池化层交叠排列用于逐层抽取NIR数据特征，输出层连接softmax分类器，对药品NIR数据进行分类概率预测。在输出层之前采用全局最大池化层，将特征图进行整体池化，形成一个特征点，用于解决全连接层存在的限制输入维度大小，参数过多的问题。同时，在网络模型中引入批处理操作和dropout机制，以防止梯度消失和减小网络过拟合的风险。在网络模型的设计过程中，通过设计不同的卷积网络层数以及不同的卷积核尺寸大小，分析其对建模效果的影响，同时分析五种经典数据预处理方法对NIR分析的影响。以我国7个厂商生产的头孢克肟片和11个厂商生产的苯妥英钠片样本NIR为实验对象，建立药品的多品种、多厂商分类模型，该模型在二分类、多分类实验中取得了良好的分类效果。在十八分类实验中，当训练集与测试集比例为7∶3时，分类准确率为99.37±0.45，比SVM，BP，AE和ELM算法取得更优的分类性能。同时，深度卷积神经网络模型推理速度较快，优于SVM和ELM算法，但训练速度慢于二者。大量实验结果表明，深度卷积神经网络可对多品种、多厂商药品NIR数据准确、可靠地判别分类，且模型具有良好的鲁棒性和可扩展性。该方法也可推广到烟草、石化等其他领域的NIR数据分类应用中。

关键词 深度卷积神经网络；近红外光谱；药品鉴别；多分类

中图分类号：TP391 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)11-3606-08

引 言

由于原材料，生产工艺以及包装形式等的差别，不同厂商生产的同一种药品的品质也有一定的差异。对这些差异性的鉴别在药品的监管过程中具有重要意义。近红外光谱(near infrared spectroscopy，NIR)分析具有高效、稳定、无损、无污染等优点，已广泛应用于食品、制药以及医疗等领域[1-4]。近红外光谱分析结合化学计量学方法已广泛应用于药品的快速、无损建模分析。Deconinck等[5]使用决策树对Viagra和Cialis两种药品的近红外光谱进行了鉴别，但该方法尚未对多分类问题进行研究。张卫东等[6]融合堆栈稀疏自编码和核极限学习机各自的优势，提出SSAE-KELM模型，对四个厂商生产的铝塑和非铝塑包装形式的头孢克肟片近红外光谱进行分类研究，取得良好的效果，但仅报道了单个药品分类研究结果。Yang等[7]结合Dropout和深度信念网络构建Dropout-DBN分类器，对琥乙红霉素药品近红外光谱进行鉴别，有效缓解了由于训练数据少导致模型过拟合现象，实验结果表明深度学习方法适合相对较小样本规模的近红外光谱数据分析。然而，随着制药厂商以及药品品种的增加，

收稿日期：2019-03-04，**修订日期**：2019-07-11

基金项目：国家自然科学基金项目(21365008，61562013)资助

作者简介：李灵巧，1986年生，北京邮电大学自动化学院博士研究生 e-mail：54pe@163.com
潘细朋，1985年生，北京邮电大学自动化学院博士研究生 e-mail：pxp201@bupt.edu.cn
李灵巧，潘细朋：并列第一作者 *通讯联系人 e-mail：fyc@nifdc.org.cn；yhh@bupt.edu.cn

药品光谱数据进一步积累，设计面向多品种、多厂商并且可扩展性良好的分类模型显得尤为重要。

深度卷积神经网络(convolutional neural network，CNN)具有极强的建模能力，广泛应用于多分类和大规模数据的计算机视觉和语音识别等领域并取得了巨大成功[8-10]。同时，由于其深层的网络结构和非线性激活能力，深度卷积网络应用到近红外光谱的建模分析已有些报道[11-12]。鲁梦瑶等[11]提出一种改进的 LeNet-5 卷积神经网络模型，对烟叶样本的近红外光谱进行分类实验研究，获得良好的分类性能。该方法本质上是将一维光谱数据转换成二维的矩阵，以适应现有的 CNN 模型。Acquarelli 等[12]设计一个含一层卷积的 CNN 模型，并用于振动光谱数据分析的分类、分析，但由于采用浅层 CNN 模型，面向大规模的光谱数据分类准确率有待提高。

尽管 CNN 网络在图像分类、语音识别等领域取得了突破性的进展，但在 NIR 分析方面的应用却比较少。其主要原因有以下两点：(1)近红外光谱数据本质上是一维向量(矩阵)，不大适合套用现有的二维 CNN 模型；(2)相比图像数据，光谱数据的获取困难的多，光谱样本量通常较小，经典的化学计量学方法可满足应用要求。但随着数据的积累，光谱类别数大幅增加，经典的分类方法无法满足高精度光谱数据鉴别的要求。基于上述分析，本文采用一维 CNN 模型用于多品种、多厂商的药品近红外光谱分类研究。核心贡献为：(1)将 CNN 模型引入药品 NIR 光谱鉴别领域；(2)基于一维 CNN，设计若干个端到端的 NIR 光谱分类模型，适用于大规模、多品种、多厂商的药品鉴别场景；(3)本方法性能良好，分类准确率超过或比肩多个现有的最佳方法，同时具有良好的鲁棒性和可扩展性，适合其他领域的 NIR 光谱数据分析。

1　算法描述

1.1　卷积神经网络

针对 NIR 光谱数据的特性，设计了多个一维的 CNN 网络模型，并进行比较研究，具体在第 1.2 节将详细介绍和讨论，本节以其中的一个 CNN 网络为例进行介绍。网络主要分为输入层、卷积层、池化层、全局最大池化层、输出层，示意图如图 1 所示。图 1 中含三个卷积层，卷积核大小分别为 21，19 和 17。卷积核的权值采用 Xavier 正态分布初始化。卷积后进行批处理化(batch normalization，BN)操作，BN 操作将每个 batch 上前一层的激活值重新标准化，将原本减小的激活值放大，防止梯度消失。池化层紧接在每个卷积层的后边，它起到减小输出大小，降低过拟合的作用。该网络采用 pool_size=3，strides=3 的池化操作。全局最大池化层的思想是将最后一层的特征图进行整体池化，形成一个特征点，主要是用来解决全连接层存在的限制输入维度大小，参数过多的问题。输出层的神经元个数为药品的类别数，通过连接 softmax 分类器，对药品 NIR 数据进行分类概率预测。softmax 公式为

$$L_{\text{softmax}} = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{k} 1\{y_i = j\}\log\frac{e^{\theta_j^{\mathrm{T}} x_i}}{\sum_{j=1}^{k} e^{\theta_j^{\mathrm{T}} x_i}}\right] \quad (1)$$

其中 $1\{y_i=j\}$ 为示性函数，当 $y_i=j$ 时，$1\{y_i=j\}=1$，否则 $1\{y_i=j\}=0$。N 是光谱样本数，k 是光谱样本类别数。θ 表示 softmax 分类器参数。

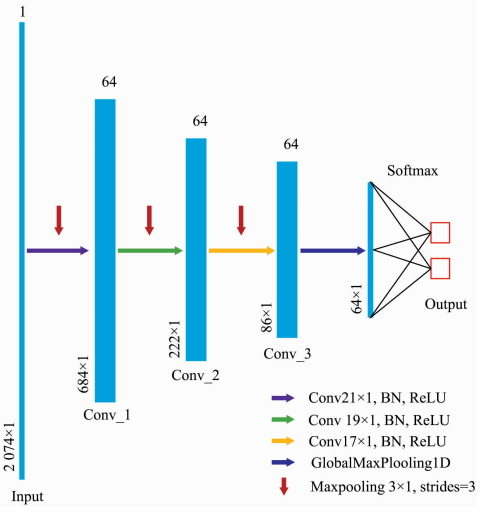


图 1　一维 CNN 模型示意图

Fig. 1　The structure of one-dimensional CNN

为减小网络过拟合的风险，在网络的每个卷积层后引入 dropout 机制[13]。为了公平对比，BP 模型和 AE 算法与本文设计的 CNN 网络模型采用同样的 dropout 值。

1.2　深度卷积神经网络的设计

为了探究卷积核大小以及网络层数对药品 NIR 分类性能的影响，设计了 7 种具有不同深度和不同卷积核大小的卷积神经网络模型，分别为 3 种大卷积核的网络，以及 4 种小卷积核的网络，取名 CNN-1—CNN-7。为节约篇幅，从大卷积核和小卷积核神经网络中选取 CNN-1 和 CNN-4 为代表进行说明，详见表 1 和表 2。其中，药品 NIR 光谱数据维度为 2074；以七分类为例，其输出层神经元个数为 7。

表 1　CNN-1 网络的各项参数

Table 1　The parameters of CNN-1

层类别	卷积核大小	卷积核数量	步长	输出形状	神经元个数	参数量
Conv_1	21	64	1	2 054×64	131 456	1 408
BN_1	—	—	—	2 054×64	131 456	256
MP_1	—	—	3	684×64	43 776	0
Conv_2	19	64	1	666×64	42 624	77 888
BN_2	—	—	—	666×64	42 624	256
MP_2	—	—	3	222×64	14 208	0
GMP	—	—	—	64	64	0
Output	—	—	—	7	7	455
合计	—	—	—	—	406 215	80 263

CNN-1 模型含有两个卷积层，卷积核大小分别为 21 和 19。卷积层后紧接着BN层以及最大池化层。第二个最大池

化层后连接全局最大池化层，最后是输出层。第一个卷积层采用 64 个尺寸为 21 的卷积核，在一维光谱上移动的步长为 1，生成 64 层特征映射图，需要训练的参数为 1 408 个。第二个卷积层采用 64 个尺寸为 19 的卷积核，生成 64 层特征映射图，需要训练的参数为77 888个。经全局最大池化层后神经元个数为 64，经全连接输出，输出为 7 分类，该层的参数为 455 个。加上 2 个 BN 层的参数，CNN-1 模型总共含有 80 263个参数。CNN-2 与 CNN-1 网络类似，比 CNN-1 多一个卷积层、BN 层以及最大池化层组合。CNN-3 比 CNN-2 同样多一个卷积层、BN 层以及最大池化层组合。

表 2 CNN-4 网络的各项参数

Table 2 The parameters of CNN-4

层类别	卷积核大小	卷积核数量	步长	输出形状	神经元个数	参数量
Conv_1	3	64	1	2 072×64	132 608	256
BN_1	—	—	—	2 072×64	132 608	256
MP_1	—	—	3	690×64	44 160	0
Conv_2	3	64	1	688×64	44 032	12 352
BN_2	—	—	—	688×64	44 032	256
MP_2	—	—	3	229×64	14 656	0
GMP	—	—	—	64	64	0
Output	—	—	—	7	7	455
合计	—	—	—	—	412 167	13 575

CNN-4—CNN-7 模型是小卷积核网络，卷积核大小都为 3，但所含卷积层数不同。CNN-4 模型与 CNN-1 模型结构类似，只是卷积核大小不同。CNN-5 与 CNN-2 模型结构类似，唯一区别是将卷积核的大小都变成 3。CNN-6 与 CNN-7 在 CNN-5 的基础上逐一增加一个卷积核、BN 层以及最大池化层组合。

1.3 卷积神经网络的训练

整个 CNN 网络模型使用 Adam 优化器进行训练，Adam 优化器由 Kingma 和 Ba 两位学者于 2014 年底提出，结合 AdaGrad 和 RMSProp 两种优化算法的优点，计算高效，收敛速度较快。模型训练过程中，使用回调函数来观察网络内部的状态和统计信息。在验证集损失在一定的周期 n(e.g.，n=10)内不下降的情况下，对学习率乘以系数 lambda(e.g.，lambda=0.2)达到减小学习率的目的。如验证集损失相比上一个训练周期没有下降，则经过 patience(e.g.，patience=50)个周期后停止训练，以防止过拟合。

1.4 其他对比实验算法简介

对比实验方法有支持向量机(support vector machines，SVM)，反向传播算法(back propagation，BP)，自编码器(autoencoder，AE)以及极限学习机(extreme learning machines，ELM)，几种方法简要介绍如下。

(1)SVM 的参数 C 设置为 1.0，gamma 设置为 0.001，其他参数使用 python 机器学习库 sklearn 的默认值。多分类实验时，采用 one-against-all 策略来构造多类 SVM 分类器。

(2)BP 算法与 CNN 模型一致，也采用三层网络结构。BP 网络结构为：2074-500-100-20-num_class(num_class 为分类类别数)。

(3)AE 模型也采用三层网络结构，AE 编码网络结构为：2074-500-100-20-num_class，解码网络结构为：num_class-20-100-500-2074。训练自编码模型时，损失函数采用均方误差损失，使用 Adam 作为优化器；当无监督预训练结束后，进行有监督分类训练。

(4)ELM 的网络结构为 2074-800-num_class。实验表明，ELM 网络隐藏层神经元个数设置为 800，模型取得较好的性能。

2 实验部分

2.1 数据采集

实验数据为中国食品药品检定研究院采集的非铝塑包装头孢克肟片和苯妥英钠片两种药品的 NIR 数据。数据通过 Bruker Matrix 光谱仪测得，每条 NIR 数据的波长范围是 4 000～11 995 cm^{-1}，间隔 4 cm^{-1}，有 2 074 个吸光度值。NIR 样品信息如表 3 和表 4 所示。7 个厂商生产的头孢克肟片样本 NIR 光谱如图 2 所示。图 3 为 11 个不同的制药厂商生产的苯妥英钠片光谱。从图 2 和图 3 中可以看出，不同厂商的同种药品光谱图非常相似，部分谱段甚至重叠，这对分类算法带来较大挑战。

表 3 头孢克肟片 NIR 数据

Table 3 NIR data of cefixime tablets

厂商	种类	数量
湖南方盛制药股份有限公司	头孢克肟片	54
江苏正大清江制药有限公司	头孢克肟片	63
山东鲁抗医药股份有限公司	头孢克肟片	51
山东罗欣药业股份有限公司	头孢克肟片	48
广州白云山制药总厂	头孢克肟片	39
四川方向药业有限责任公司	头孢克肟片	48
苏州东瑞制药有限公司	头孢克肟片	30
	共计	333

表 4 苯妥英钠片 NIR 数据

Table 4 NIR data of phenytoin tablets

厂商	种类	数量
东北制药集团沈阳第一制药厂	苯妥英钠片	42
河北东风药业有限公司	苯妥英钠片	30
开封制药(集团)有限公司	苯妥英钠片	30
山西汾河制药有限公司	苯妥英钠片	47
山西省临汾健民制药厂	苯妥英钠片	72
山西云鹏制药有限公司	苯妥英钠片	135
上海信谊黄河制药有限公司	苯妥英钠片	87
石药集团欧意药业有限公司	苯妥英钠片	50
天津力生制药股份有限公司	苯妥英钠片	98
西南药业股份有限公司	苯妥英钠片	71
扬州市星斗药业有限公司	苯妥英钠片	54
	共计	716

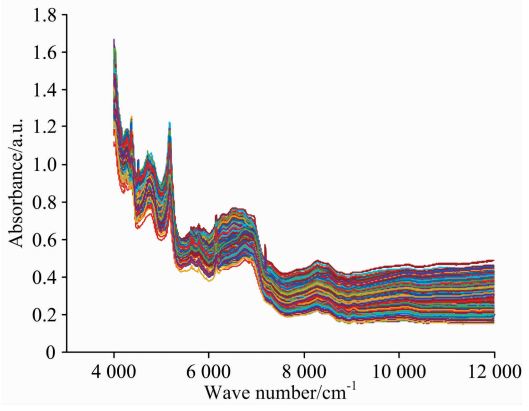


图 2　7 个制药厂商头孢克肟片光谱曲线图

Fig. 2　NIR of cefixime tablets from seven pharmaceutical manufacturers

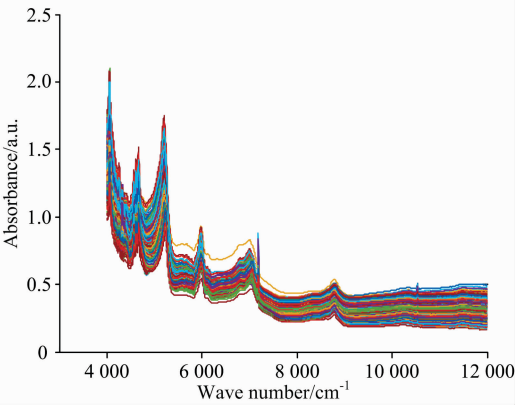


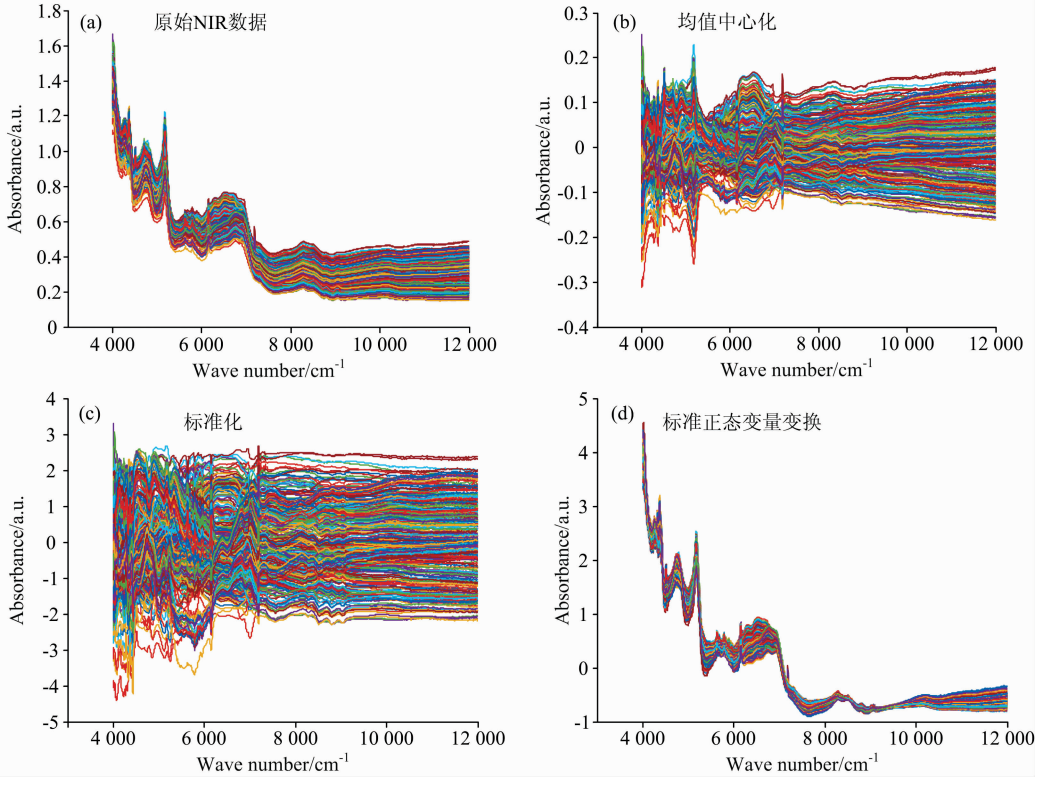
图 3　11 个制药厂商苯妥英钠片光谱曲线图

Fig. 3　NIR of phenytoin tablets from eleven pharmaceutical manufacturers

2.2　光谱数据预处理

采用均值中心化、标准化、标准正态变量变换、Savitzky-Golay(SG)平滑求导以及多元散射校正 5 种经典方法对 NIR 光谱进行预处理。以 7 个制药厂商生产的头孢克肟片光谱为例，分析 5 种光谱数据预处理方法的效果。从图 4 可以看出，均值中心化方法可以使光谱之间的差异性增大，标准化方法能有效克服光谱数据中存在的噪声点和异常值，SG 算法对 NIR 光谱曲线进行平滑，可有效消除光谱中噪声。

对每一个光谱预处理方法处理后的数据进行分类实验，选取性能最佳的预处理方法，实验结果详见表 5。表 5 中，ratio 表示训练集占总数据的比率，acc_raw，acc_cen，acc_auto，acc_snv，acc_savg 和 acc_msc 分别代表 NIR 数据未经预处理以及经过上述 5 种预处理测试集的分类准确率。随机重复 5 次实验取均值和标准差作为实验结果，实验中迭代周期设为 100。从表 5 的实验结果可以看出，相比原始光谱实验结果，均值中心化、标准化、标准正态变量变换后的分类准确率有较大的提高，经过标准化后的光谱数据在大多数情况下取得最佳分类结果。因此，后续的分类实验中选取标准化为预处理方法。



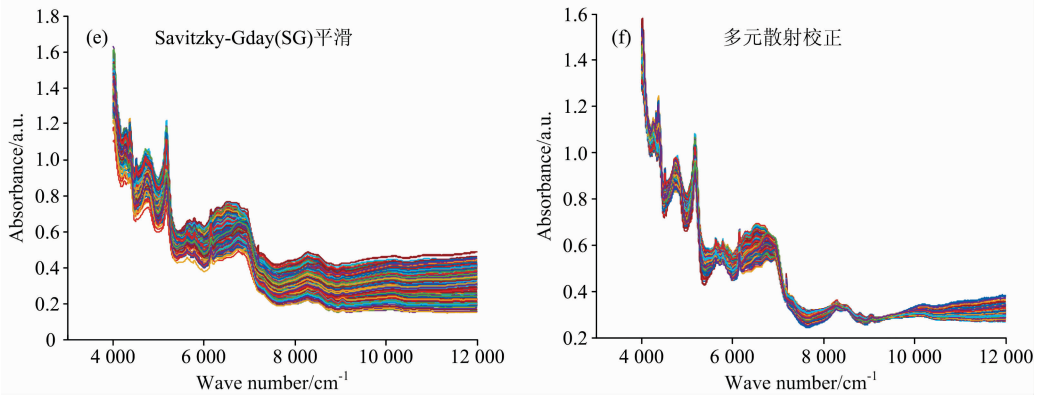


图 4 光谱原图与 5 种方法预处理后光谱曲线图

Fig. 4 The original NIR and the NIR preprocessed by five methods

2.3 建立校正模型

以 7 个厂商头孢克肟片未经预处理的 NIR 数据为例，通过对每一个 CNN 模型重复多次(5 次)实验，计算光谱数据分类准确率，选取性能最优的模型，用于光谱数据分类实验，实验结果如表 6 所示。

表 6 中，ratio 的含义与表 5 一致。acc_cnn1—acc_cnn7 分别代表上述七种 CNN 网络模型在测试集上的分类准确率。从表 6 中可以看出，CNN-1—CNN-3 这 3 个模型获得较优的分类性能，而 CNN-4—CNN-7 这 4 个模型性能较差。主要原因是 CNN-4—CNN-7 模型采用小的卷积核，难以捕获光谱曲线有区分度的模式特征，而 CNN-1—CNN-3 中的卷积核尺寸较大，在光谱曲线上滑动时跨度较大，能有效提取连续变化分辨率高的光谱曲线的模式特征。同时，我们观察到 CNN-2 比 CNN-1 和 CNN-3 整体性能更优。CNN-1—CNN-3 分别包含 2～4 个卷积层。对于样本量较小的光谱数据，采用层数过多的模型，容易出现过拟合，但浅层网络建模能力较弱，综合考虑，采用 CNN-2 模型进行后续实验研究。

表 5 不同预处理 NIR 数据在 CNN-2 模型的分类性能

Table 5 Classification performance of NIR data with different preprocessing methods in CNN-2 model

ratio	acc_raw	acc_cen	acc_auto	acc_snv	acc_savg	acc_msc
0.8	87.35±6.21	99.41±0.72	99.12±0.72	99.71±0.59	84.71±12.43	79.71±11.89
0.7	85.74±9.51	98.02±0.0	99.01±0.89	97.23±3.22	81.58±8.62	69.5±15.68
0.6	79.7±10.87	97.14±1.46	99.1±0.56	99.25±0.48	79.55±3.76	52.33±14.89
0.5	79.03±5.67	98.67±0.8	99.03±0.91	97.45±1.85	71.52±11.08	65.82±7.44
0.4	77.59±5.99	97.59±0.86	98.39±0.74	91.96±6.85	70.65±13.95	68.04±19.7
0.3	58.53±9.49	97.49±1.07	98.18±0.69	97.49±1.63	69.35±8.21	51.77±15.64
0.2	60.45±4.32	94.39±2.01	95.53±2.25	82.12±10.61	62.27±4.1	63.71±10.96

表 6 七种 CNN 模型的分类性能对比

Table 6 Comparison of classification performance of seven CNN models

ratio	acc_cnn1	acc_cnn2	acc_cnn3	acc_cnn4	acc_cnn5	acc_cnn6	acc_cnn7
0.8	70.29±15.15	85.0±7.11	74.41±8.95	18.24±1.5	28.82±4.32	25.29±5.84	37.06±9.08
0.7	77.03±14.31	77.03±10.32	68.12±11.09	18.22±3.35	32.67±2.58	22.77±6.2	31.49±6.78
0.6	79.1±15.06	69.62±4.53	65.26±6.86	18.8±1.96	32.03±3.58	23.01±4.13	27.82±6.4
0.5	58.18±14.05	70.67±9.41	57.45±8.2	18.79±4.0	22.18±4.55	22.55±6.3	24.97±4.12
0.4	53.37±14.31	75.38±3.05	55.58±5.23	17.19±0.97	19.6±3.85	21.71±5.12	18.29±4.27
0.3	44.68±18.42	67.79±6.74	47.79±6.39	20.78±7.3	21.73±5.55	20.61±6.25	19.48±9.13
0.2	35.3±15.42	56.97±8.07	47.88±5.33	18.79±1.28	20.0±8.14	20.23±5.25	14.85±2.79

实验代码采用 python2.7 编写，硬件环境 GPU 型号为 NVIDIA Tesla P100。在二分类和多分类光谱数据上进行实验，并与当前最佳 NIR 光谱分类算法，如 SVM，BP，AE 以及 ELM 进行比较。

3　结果与讨论

本工作的实验设置思路为：首先在头孢克肟片 NIR 数据上分别进行二分类和七分类实验，然后结合苯妥英钠片 NIR 数据进行多药品多厂商的十八分类实验。

3.1　二分类实验

参照文献[6]的实验数据设置进行二分类实验。实验样本设置为：取表 3 中江苏正大生产的头孢克肟片 NIR 样本共 63 个，作为负类样本集；取湖南方盛、山东鲁抗和山东罗欣三个厂商生产的头孢克肟片 NIR 样本共 153 个，作为正类样本集。两组数据的实验结果如表 7 所示。二分类实验由于分类性能都较好，不能体现本文设计的 CNN 模型的优越性。

表 7　二分类实验不同方法分类准确率

Table 7　Classification accuracy of different methods in binary classification experiment

ratio	acc_cnn	acc_svm	acc_bp	acc_ae	acc_elm
0.8	100.0±0.0	100.0±0.0	100.0±0.0	94.09±11.82	100.0±0.0
0.7	100.0±0.0	100.0±0.0	100.0±0.0	94.15±11.69	100.0±0.0
0.6	100.0±0.0	99.77±0.47	98.84±1.8	99.3±0.93	100.0±0.0
0.5	99.81±0.37	99.81±0.37	99.25±0.7	92.9±11.07	100.0±0.0
0.4	99.22±0.98	99.69±0.38	99.69±0.62	99.22±1.2	100.0±0.0
0.3	99.07±0.53	99.33±0.42	98.4±0.68	87.73±13.15	99.07±0.68
0.2	97.78±1.19	98.71±0.86	99.06±0.7	91.7±10.69	99.18±0.79

3.2　多分类实验

(1)七分类实验。表 3 中 7 个厂商生产的头孢克肟片 NIR 样本分别作为不同的类，计 7 类。实验结果如表 8 所示，CNN 在 ratio=0.7，0.6 时，相比其他方法，获得最高分类准确率；ELM 算法在 ratio 较小时，获得最佳性能，但在这些情况下，CNN 与 ELM 性能相近。相同条件下，CNN 比 SVM 高 1～3 个点，比 BP 高 1～2.8 个点，比 AE 高 4～21 个点。图 5 为七分类 ratio=0.7 时，迭代周期为 100 的 ROC 曲线，从图中可以看出，ELM 方法的 AUC 值为 1，CNN 的 AUC 值为 0.9997，接近 1，说明两个方法取得非常好的分类性能。SVM，BP 和 AE 方法的 AUC 值也大于 0.9，表明分类性能也不错。该结论从表 8 中也可以得到佐证。

表 8　七分类实验不同方法分类准确率

Table 8　Classification accuracy of different methods in seven classification experiments

ratio	acc_cnn	acc_svm	acc_bp	acc_ae	acc_elm
0.8	99.12±1.18	98.53±1.61	97.94±1.18	95.29±5.98	99.41±1.18
0.7	98.81±1.15	95.64±1.34	97.82±1.92	86.73±8.18	98.61±1.34
0.6	99.25±0.67	97.74±1.26	98.65±1.67	77.74±20.23	98.95±0.77
0.5	98.3±0.89	96.97±1.15	97.21±0.98	77.82±15.93	98.91±0.8
0.4	96.18±2.47	93.67±2.87	95.08±1.57	79.1±12.87	97.89±0.86
0.3	96.1±2.27	93.25±1.59	93.59±2.42	72.03±7.59	97.66±1.05
0.2	93.94±1.1	91.06±1.6	91.14±3.21	78.26±10.17	97.42±1.08

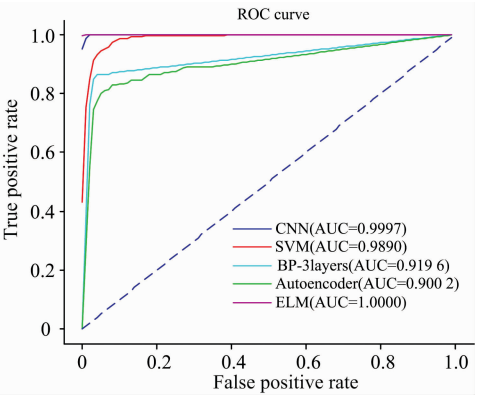


图 5　ratio=0.7 时七分类实验 ROC 曲线

Fig. 5　ROC curve of seven classification experiment when ratio is 0.7

(2)十八分类实验。取表 3 中 7 个厂商生产的头孢克肟片光谱样本以及表 4 中 11 个不同厂商生产的苯妥英钠片 NIR 数据分别作为不同的类，共 18 类。实验结果如表 9、表 10 所示。表 9 是十八分类实验不同方法分类准确率，环比二分类和七分类实验结果，CNN 方法在十八分类实验中仍然获得了非常好的性能，但其他方法性能下降幅度较大，尤其是 SVM 和 ELM，体现了 CNN 方法对于多药品、多厂商分类的优越性。实验中采用 one-against-all 策略构造多类 SVM 分类器，由于算法策略带来的样本不平衡问题，导致分类准确率一定程度的降低。ELM 模型在隐藏层的权重和偏置被随机确定后，隐藏层的输出矩阵 $\boldsymbol{H}$ 就被唯一确定。训练单隐层神经网络可以转化为求解一个线性方程 $\boldsymbol{H}\beta=\boldsymbol{T}$，并且输出权重 β 可以被确定为 $\hat{\beta}=\boldsymbol{H}^{*}\boldsymbol{T}$。然而，ELM 这种模型的求解方式容易出现过拟合的现象。表 10 是十八分类实验各方法的训练和推理时间列表。从表中可以看出，所有算法随着训练集所占比率的减小，训练时间呈下降趋势。SVM 算法的模型训练时间最少，ELM 算法次之，BP，AE 以及 CNN 训练时间开销较大。推理时间方面，BP，AE 以及 CNN 在一个数量级，速度较快，而 SVM 和 ELM 的推理时间比其他 3 种方法多一个数量级。从时间维度来看，模型推理时间往往更值得关注，因为模型一般是离线训练，然后再上线部署。图 6 是十八分类 ratio=0.2 时，迭代周期为 100 的 ROC 曲线图。

表 9　十八分类实验不同方法分类准确率

Table 9　Classification accuracy of different methods in eighteen classification experiment

ratio	acc_cnn	acc_svm	acc_bp	acc_ae	acc_elm
0.8	99.16±0.75	88.69±2.22	90.84±1.86	89.07±6.52	80.84±3.4
0.7	99.37±0.45	87.38±1.0	90.73±2.86	95.33±1.87	78.55±3.9
0.6	98.0±0.32	86.9±1.29	90.71±1.78	91.1±4.31	86.1±1.19
0.5	97.81±0.6	84.8±1.08	86.76±4.25	88.06±7.01	88.25±0.87
0.4	96.17±0.23	82.71±2.14	85.23±5.76	83.41±4.63	90.3±1.22
0.3	94.55±1.61	80.41±1.73	83.86±3.31	80.41±4.84	88.22±2.28
0.2	89.03±1.92	76.59±2.48	81.2±2.86	74.48±3.75	85.88±1.66

表 10 十八分类实验不同方法训练时间和推理时间(单位：秒)

Table 10 Training and inference time of different methods in eighteen classification experiments (unit: s)

ratio	ttime_cnn	ttime_svm	ttime_bp	ttime_ae	ttime_elm
0.8	88.13	3.37	28.52	27.91	2.29
0.7	80.03	2.7	27.91	27.2	1.94
0.6	73.06	2.16	27.09	26.47	1.45
0.5	64.41	1.67	26.08	24.51	1.14
0.4	56.91	1.16	24.62	23.96	0.75
0.3	50.68	0.71	24.52	22.02	0.5
0.2	26.13	0.39	23.26	23.28	0.28
ratio	itime_cnn	itime_svm	itime_bp	itime_ae	itime_elm
0.8	0.04	0.41	0.01	0.01	0.1
0.7	0.05	0.56	0.01	0.01	0.18
0.6	0.07	0.67	0.02	0.02	0.28
0.5	0.09	0.77	0.03	0.03	0.41
0.4	0.06	0.79	0.03	0.03	0.62
0.3	0.1	0.76	0.03	0.04	0.85
0.2	0.08	0.68	0.05	0.04	1.09

5 种方法获得的 AUC 值都较大，证明这些方法都较适合 NIR 光谱数据的分类，其中，CNN 模型获得最大 AUC 值，从另一个角度也证明了 CNN 模型在多分类准确率方面的优越性。

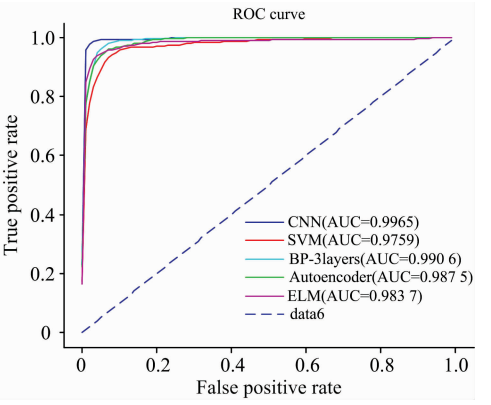


图 6 ratio=0.2 时十八分类 ROC 曲线

Fig. 6 ROC curve of eighteen classification experiment when ratiois 0.2

4 结 论

设计了几种简单但非常有效的一维深度卷积网络模型，用于药品的 NIR 鉴别分析，并对卷积网络层数以及卷积核尺寸对建模结果的影响进行详细实验研究，同时分析五种经典预处理方法对药品 NIR 数据分析的影响。与当前最佳方法，比如：SVM，BP，AE，ELM 算法进行对比，在大规模，多药品、多厂商 NIR 鉴别实验中取得了更高的分类准确率和良好的可扩展性。一维深度卷积网络模型推理阶段速度较快，优于 SVM 和 ELM 算法，但训练速度慢于二者。

References

[1] Ma H L, Wang J W, Chen Y J, et al. Food Chemistry, 2017, 215: 108.

[2] Lê L M M, Eveleigh L, Hasnaoui I, et al. Journal of Pharmaceutical and Biomedical Analysis, 2017, 138: 249.

[3] Xue J T, Ye L M, Li C Y, et al. Optik, 2018, 170: 30.

[4] Risoluti R, Materazzi S, Gregori A, et al. Talanta, 2016, 153: 407.

[5] Deconinck E, Sacré P Y, Coomans D, et al. Journal of Pharmaceutical and Biomedical Analysis, 2012, 57: 68.

[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] Yang H H, Hu B C, Pan X P, et al. Journal of Innovative Optical Health Sciences, 2016, 10(2): 1630011.

[8] Lecun Y, Bengio Y, Hinton G. Nature, 2015, 521(7553): 436.

[9] Nassif A B, Shahin I, Attili I, et al. IEEE Access, 2019, 7: 19143.

[10] Lai D, Tian W, Chen L. Pattern Recognition, 2019, 88: 547.

[11] LU Meng-yao, YANG Kai, SONG Peng-fei, et al(鲁梦瑶，杨 凯，宋鹏飞，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(12): 3724.

[12] Acquarelli J, Laarhoven T V, Gerretzen J, et al. Analytica Chimica Acta, 2017, 954: 22.

[13] Srivastava N, Hinton G, Krizhevsky A, et al. Journal of Machine Learning Research, 2014, 15(1): 1929.

Deep Convolution Network Application in Identification of Multi-Variety and Multi-Manufacturer Pharmaceutical

LI Ling-qiao[1,2], PAN Xi-peng[1], FENG Yan-chun[3*], YIN Li-hui[3], HU Chang-qin[3], YANG Hui-hua[1,2*]

1. School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
3. National Institutes for Food and Drug Control, Beijing 100050, China

Abstract As near infrared spectroscopy (NIR) has many advantages, such as high efficiency, being non-destructive and environment-friendly and on-site detection, it is especially suitable for rapid modeling and analysis of drugs. However, there are some shortcomings such as weak absorption intensity and overlapping bands. It is necessary to establish a robust and reliable chemometrics model to analyze NIR. Deep convolution neural network (DCNN) is an important branch of deep learning method, which extracts data features layer by layer, combines and transforms them to form higher-level semantic features. It is widely used in computer vision, speech recognition and other fields, and has achieved great success, but has not been reported in drug NIR analysis yet. Based on the deep convolution network model, this paper studies the multi-class modeling of drug NIR. According to the characteristics of drug NIR data, several one-dimensional deep convolution network models for multi-class and multi-manufacturer drug NIR classification are designed. The overlapping arrangement of convolution layer and pool layer in the model is employed to extract NIR data features layer by layer, and the output layer is connected with the softmax classifier to predict the classification probability of NIR data. Before the output layer, the global maximum pooling layer is used to solve the problem of restricting the size of input dimension and too many parameters in the full connection layer. At the same time, batch normalization and dropout are introduced in the network model to prevent the gradient vanishing and reduce the risk of network overfitting. The impact on the modeling effect with different convolutional network layers and different convolution kernel sizes is analyzed. At the same time, the influence of five classical data preprocessing methods is explored. Taking NIR samples of cefixime and phenytoin tablets as experimental datasets, a multi-class and multi-manufacturer classification model of drugs is established. The model achieved good classification results in the experiments of binary-classification and multi-classification. In eighteen classification experiments, when the ratio between training set and test set was 7∶3, the classification accuracy was 99.37±0.45, which achieved better classification performance than SVM, BP, AE and ELM. At the same time, inference speed of deep convolution neural network was faster than SVM and ELM, but training speed was slower than both. A large number of experimental results showed that the deep convolutional neural network can accurately and reliably distinguish the NIR data of multi-class and multi-manufacturer drugs, with good robustness and scalability. The proposed method can also be extended to the application of NIR data classification in tobacco, petrochemical and other fields.

Keywords Deep convolution neural network; Near infrared spectroscopy; Pharmaceutical discrimination; Multi-classification

(Received Mar. 4, 2019; accepted Jul. 11, 2019)

LI Ling-qiao and PAN Xi-peng: joint first authors

* Corresponding authors