基于堆栈压缩自编码的近红外光谱药品鉴别方法

甘博瑞[1]，杨辉华[1,2*]，张卫东[1]，冯艳春[3]，尹利辉[3]，胡昌勤[3]

1. 桂林电子科技大学电子工程与自动化学院，广西 桂林 541004
2. 北京邮电大学自动化学院，北京 100876
3. 中国食品药品检定研究院，北京 100050

摘 要 由于近红外光谱在药品鉴别应用中具有分析速度快、样品无损、可现场检测等突出优点，目前已在众多领域中广泛应用。但近红外光谱存在信噪比低，吸收强度弱且谱峰重叠等缺点，无法从光谱中直接得到定性/定量的物质信息，因而近红外光谱分析技术常作为一种间接分析技术，并且光谱的化学计量学建模方法成为近红外光谱分析的核心内容。深度学习是机器学习的一个新的分支，并已经成功运用于多个领域。深度学习的网络结构和非线性的激活能力，使其模型特别适合高维、非线性的大规模数据建模。为进一步丰富近红外光谱建模方法，并提高近红外光谱分析技术的回归精度或分类准确率，将深度学习方法应用于近红外光谱分析，发展新的建模方法十分必要。面向近红外光谱定性分析技术，提出一种基于堆栈压缩自编码网络(SCAE)光谱定性分析方法，并应用于多类别药品的光谱分析，以区分或鉴别不同厂家生产的同种药品。压缩自编码网络(CAE)以自编码网络(AE)为基础，进一步加入雅克比矩阵作为约束项。自编码网络最初是用实现数据降维，以学习数据内部特征，而雅克比矩阵包含数据在各个方向上的信息，将其作为AE的约束项则可使提取到的特征对输入数据在一定程度下的扰动具有不变性，从而提高AE提取特征的能力。SCAE是一种由多层CAE构成的神经网络。前一层CAE的隐藏层作为后一层CAE的输入层，网络的全部参数是通过采用逐层贪婪的训练方式来获取的，训练结束后将所有网络视为一个整体，通过反向传播算法进行微调，最后使用Logistic/Softmax分类器进行定性分析。实验数据均为中国食品药品检定研究院采集，以头孢克肟胶囊作为二分类实验数据，硝酸异山梨酯片作为多分类实验数据。通过Bruker Matrix光谱仪测定每个样本在不同波长下的吸光度值得到其光谱曲线，再通过OPUS软件消除漂移等因素对光谱样本之间产生的偏差。接下来通过实验确定约束项雅克比矩阵的系数 λ 为0.003之后建立模型。建模过程分为五个阶段，分别为：预处理阶段，预训练阶段，微调阶段，测试阶段和对比阶段。为了验证SCAE在分类准确性、算法稳定性和建模时间等方面的性能，与BP神经网络、SVM算法、稀疏自编码(SAE)和降噪自编码(DAE)开展对比实验研究。分类准确性方面，在不同的训练集与测试集的比例下，SCAE均有最佳的分类准确性与算法稳定性。建模时间方面，由于SVM算法不需要预训练和特征提取，所以运行时间方面比其他算法有大的优势，但是SCAE建模速度优于除SVM之外的其他对比算法。综合而言，使用SCAE进行药品鉴别有效可行。

关键词 堆栈压缩自编码；雅克比矩阵；近红外光谱；药品鉴别

中图分类号：TP391 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)01-0096-07

引 言

假劣质药给人民生活带来了巨大的危害，检验药品真伪是一项重要的监管任务。同时，药品的生产方式、包装、成分存在差异，甚至同一药品不同厂商生产、或不同规格，也往往存在着差异，如何鉴别这种差异性，成为药品监督的主要问题。传统的检测通过化学方法分析药品成分，然后将分析结果反馈给药厂，由于分析速度慢，反馈不及时等问题，降低了样品的生效率。近红外光谱仪具有体积小，分析速度快以及受到外界因素影响小等优点，可以方便地安装在药品生产流水线上，在短时间内监控整条生产流水线，直接非破

收稿日期：2017-12-07，**修订日期**：2018-04-11

基金项目：国家自然科学基金项目(21365008，61562013)资助

作者简介：甘博瑞，1990年生，桂林电子科技大学电子工程与自动化学院硕士研究生 e-mail：276685422@qq.com

*通讯联系人 e-mail：13718680586@139.com

坏性的监控加工过程中的药品并及时的发现问题，从而保证一整批药剂的质量。由于其快速、不破坏样本、不污染环境等优势，近红外光谱分析技术广泛应用于药品检测[1]、农业产品的质量检测、食品工业、石油化工等领域。

目前已有一些近红外光谱分析法和化学计量学、机器学习方法相结合的技术应用于光谱无损快速类别分析。如，Fontalvo-GóMez M 等[2]采用近红外光谱结合主成分分析(principal component analysis，PCA)方法，在近红外检测领域取得了良好的效果。Deconinck 等[3]运用决策树建立分类模型，对于 Viagra 和 Cialis 药品的光谱进行分类，取得了较好的分类准确率。Zou 等[4]使用 SVM 算法建立分类模型，对淀粉近红外光谱进行分类实验，实验结果表明 SVM 算法比决策树具有更高的分类准确性。国内，樊书祥等采用最小二乘支持向量机(LS-SVM)建立分类模型，对梨可溶性固形物含量进行分类，在减少了计算时间的同时，取得了比 SVM 算法更高的分类准确率。在理论方面，陆婉珍、梁逸曾等在定性分析和定量分析做出了许多创新。

深度学习[5]是一种基于对数据进行表征学习的方法，其通过非监督或半监督的方式学习数据的特征来取代人工获取特征。深度学习是机器学习的一个新的分支，并成功运用在图像识别[6]、自然语言处理、机器翻译[7]等领域。同时，由于深度学习的网络结构和非线性的激活能力，使其模型特别适合高维、非线性的大规模数据建模[8]。自编码网络属于深度学习中的一种，具有学习数据内部特征的能力。将多个自编码网络连接起来，在最后加入分类器形成的模型可以用来进行分类。传统的自编码模型如稀疏自编码、降噪自编码等在分类准确率和运行时间上都存在可以提高的地方。本文针对药品鉴别问题，提出了一种基于堆栈压缩自编码[9](stacked contractive auto-encoders，SCAE)的方法。该算法通过加入雅克比矩阵作为损失函数的约束项，解决了自编码网络中过度学习的问题，既提高了分类准确性，又减少了运行时间。为验证算法有效性，通过对头孢克肟胶囊进行真假药鉴别，和对硝酸异山梨酯片的多分类实验，验证 SCAE 的算法准确性，算法稳定度和算法运行时间。并且与 BP 神经网络、SVM 算法、稀疏自编码(SAE)和降噪自编码(DAE)进行比较，结果表明 SCAE 算法在药品近红外光谱的鉴别中更加有效。

1　算法描述

1.1　自编码网络

自编码网络是一种特殊的神经网络，它由三层网络组成，包括输入层，隐藏层和输出层，如图 1 所示。

它是一种无监督的学习算法，其网络结构中输入层与输出层具有相同的神经元个数，通过反向传播算法来微调，使得输出数据与输入数据尽可能相等，从而学习到数据内部的特征。

自编码网络模型首先将输入数据 x 通过映射函数 s_f 输出到隐藏层 h，这一步为编码(encoder)过程。然后隐藏层的数据 h 通过映射函数 s_g 重构到输出层 y，这一步称为解码

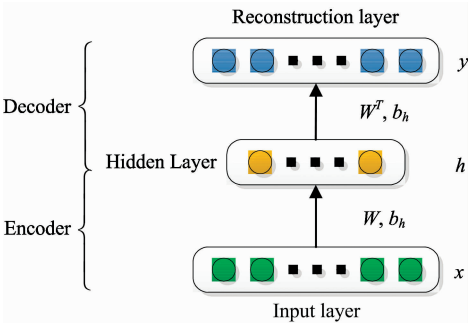

图 1　自编码结构

Fig. 1　The structure of auto-encoder

(decoder)过程。以上两个步骤可以用式(1)和式(2)表示

$$h = f(x) = s_f(Wx + b_h) \tag{1}$$

$$y = g(h) = s_g(W'h + b_y) \tag{2}$$

其中，W 是连接输入层到隐藏层的权重矩阵，W' 是 W 的转置；b_h 为隐藏层的偏置，b_y 为输出层的偏置。s_f 和 s_g 都是非线性激活函数，通常情况下可以选择 $\text{sigmoid}(z)=\frac{1}{1+e^{-z}}$ 函数或者 $\tanh(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$ 函数。本文使用 sigmoid 函数作为激活函数。

自编码网络的目的是尝试逼近一个恒等函数，使得输出值尽可能接近输入值，学习的过程就是不断地减小输出值与输入值之间的重构误差，表达式如式(3)

$$J_{\text{AE}}(\theta) = \sum_{x\in D_n} L\{x, g[f(x)]\} \tag{3}$$

式(3)中，L 为重构误差，一般情况下 L 有两种选择，可以选择均方误差 $L(x, y)=\| x-y \|^2$，或者选择交叉熵损失函数 $L(x, y) = -\sum_{i=1}^{d_x} x_i \log(y_i) + (1-x_i)\log(1-y_i)$，由于实验所用的网络结构不深，不会出现梯度消失等问题，所以最终选择均方误差作为实验的重构误差。自编码网络的学习过程就是通过反向传播算法不断减小重构误差，优化参数 W，b_h，b_y。通常情况下自编码网络中隐藏层的神经元数量要少于输入层，这相当于对输入数据进行了压缩，提取了输入数据中隐含的一些特征。事实上，自编码网络学习得到的输入数据的低维表示与主成分分析(PCA)的结果非常相似。

1.2　压缩自编码网络

压缩自编码(contractive auto-encoder)是自编码的一个变种，简称 CAE。它是在自编码的损失函数上加入了一个约束项。通常的情况下，一般自编码对权值进行惩罚的数学表达式为式(4)

$$J_{\text{AE+wd}}(\theta) = \left\{\sum_{x\in D_n} L[x, g(f(x))]\right\} + \lambda \sum_{ij} W_{ij}^2 \tag{4}$$

这是直接对 W 的值进行惩罚的表达式。CAE 的数学表达式与自编码不同，如式(5)

$$J_{\text{CAE}}(\theta) = \sum_{x\in D_n} \{L[x, g(f(x))]\} + \lambda \| J_f(x) \|_F^2 \tag{5}$$

其中 $J_f(x)$ 是隐藏层输出值关于权重的雅克比矩阵，λ 为权衡损失函数和约束项之间的比例系数，λ 的取值会在实验中

确定。而 $\| J_f(x) \|_F^2$ 表示的是雅克比矩阵的 F 范数的平方，即雅克比矩阵中每个元素求平方再求和，见式(6)

$$\| J_f(x) \|_F^2 = \sum_{ij} \left(\frac{\partial h_j(x)}{\partial x_i} \right)^2 \tag{6}$$

雅克比矩阵的 F 范数的平方求和可以写成更加具体的数学表达式，见式(7)

$$\| J_f(x) \|_F^2 = \sum_{i=1}^{d_h} [h_i(1-h_i)]^2 \sum_{j=1}^{d_x} W_{ij}^2 \tag{7}$$

式(7)中，h_i 为隐藏层的输出，w_{ij} 为输入层与隐藏层的连接权重。引入雅克比矩阵相当于对输入数据做了一个类似升维的操作，再经过特征编码之后获得了原始输入空间下的高维流形。通过计算局部流形的一阶导数使得在高维流形上的每一个点具有局部不变性，于是产生了雅克比矩阵。通过在自编码网络中引入雅克比矩阵的 F 范数作为约束项来促使学到的特征具有局部不变性。自编码网络最初是用来给数据降维，学习数据内部特征的，而雅克比矩阵包含数据在各个方向上的信息，可以使得提取到的特征对输入数据在一定程度下的扰动具有不变性。总之，压缩自编码算法主要抑制了输入数据在所有方向上的扰动。

1.3 堆栈压缩自编码网络

堆栈压缩自编码是一种由多层压缩自编码构成的神经网络。前一层压缩自编码网络的隐藏层作为后一层压缩自编码网络的输入层。如图 2 所示。

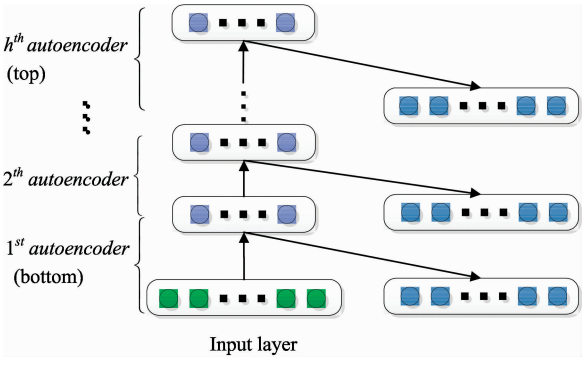

图 2 堆栈压缩自编码结构

Fig. 2 The structure of stacked contractive auto-encoders

堆栈压缩自编码网络的全部参数是通过采用逐层贪婪的训练方式来获取的，即首先训练第一个压缩自编码网络，当满足停止训练的要求时，记录下第一个压缩自编码网络中连接输入层到隐藏层的权重矩阵和偏置向量，再将训练好的第一个压缩自编码网络的隐藏层作为第二个压缩自编码网络的输入层继续训练，训练停止后将得到第二层的权重矩阵和偏置向量；以此类推可以得到堆栈压缩自编码网络的全部参数。

1.4 分类器

堆栈压缩自编码网络训练完成后，在最后一层压缩自编码网络之后接入分类器，二分类实验使用 Logistic 分类器，函数形式如式(8)

$$h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^T x}} \tag{8}$$

训练阶段，Logistic 分类器和堆栈压缩自编码网络看作一个整体，通过反向传播算法更新参数，当满足停止训练的条件时，得到最终参数 θ。测试阶段，将一条药品光谱数据传入模型，数据最后通过 Logistics 分类器输出一个结果，从函数结构可以看出其取值范围在(0，1)之间，所以可以把输出的结果看作是一种概率，当结果大于设定的阈值时，预测为一类，反之则预测为另一类。

多分类实验使用 Softmax 分类器，函数形式如式(9)

$$f(z)_j = \frac{\mathrm{e}^{z_j}}{\sum_{k=1}^{K} \mathrm{e}^{z_k}} \tag{9}$$

式(9)中，f 为映射函数，K 为样本种类个数。Softmax 分类器和 Logistics 分类器类似，训练阶段同样通过反向传播算法获得参数，不同的地方在于测试阶段输入一条药品光谱数据，最后将输出该数据属于各个类别的概率。最终选取概率最大的类别作为预测结果。

1.5 训练与测试

训练阶段将堆栈压缩自编码网络看成一个整体通过反向传播开始微调。微调是深度学习中的常用策略。从更高的视角来看，逐层训练得到的堆栈自编码网络在微调阶段被视为一个整体，这样每迭代一次，网络中所有的参数都将被优化，从而可以大幅度提升堆栈自编码网络的性能。通过多次的前向传播和反向传播过程，神经元之间的参数得到最优值，当输出结果与实际结果的误差满足要求或达到最大迭代次数时，停止学习。最后通过测试样本对模型的分类准确性进行测试。

2 实验部分

2.1 数据

实验数据均为中国食品药品检定研究院采集，其中样本集 A 包括广州白云山制药总厂以及其他药厂生产的头孢克肟胶囊，共计 252 个样本，用于二分类实验；样本集 B 包括山西云鹏制药有限公司以及其他药厂生产的硝酸异山梨酯片，共计 314 个样本，用于多分类实验。通过 Bruker Matrix 光谱仪测定每个样本在不同波长下的吸光度值得到其光谱曲线，其中每个光谱数据的波长范围是 4 000～11 995 cm^{-1}，间隔 4 cm^{-1}。样本的来源情况如表 1 所示。

表 1 药品样本概况

Table 1 The profile of the pharmaceutical samples

Dataset	厂商	数量
A	天津华津制药厂	66
	天津医药集团津康制药	90
	广州白云山制药总厂	96
总计		252
B	南京白敬宇制药有限震任公司	76
	山西云鹏制药有限公司	89
	太原市振兴制药有限公司	74
	天津太平洋制药有限公司	75
总计		314

2.2　数据预处理

图 3(a)和(b)所示分别为样本 A 和样本 B 的原始光谱图，从图中可以看出原始光谱样本重叠严重，信息解析复杂。因此首先通过 OPUS 软件消除漂移等因素对光谱样本之间产生的偏差。之后再对数据进行归一化操作，将各个波长点的吸度光度值控制在[0，1]之间，归一化的作用是用来消除光谱之间数量级的差别，使得寻找最优解的过程会变得平缓，更容易收敛到最优解。

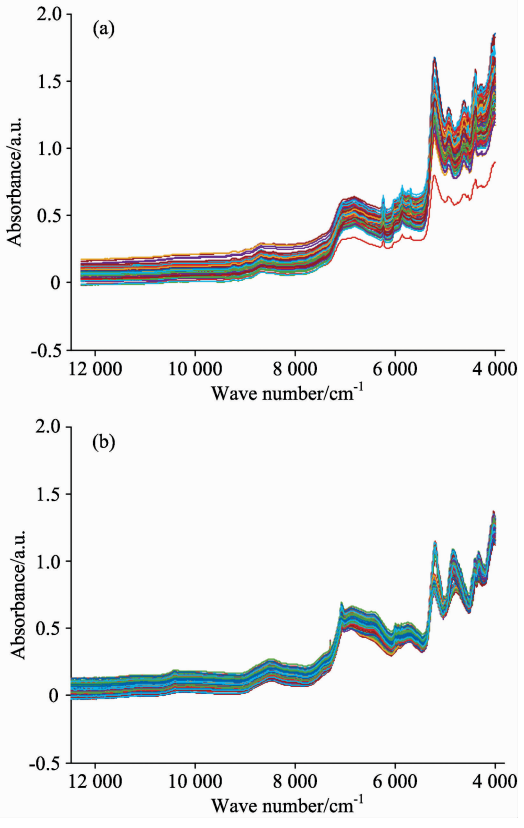

图 3　原始样本 A(a)和 B(b)的近红外光谱图

(a)：样本 A 原始光谱图；(b)：样本 B 原始光谱图

Fig. 3　The NIR spectra (a, b) of pharmaceutical samples (A, B)

(a): The NIR spectra of pharmaceutical samples A;
(b): The NIR spectra of pharmaceutical samples B

2.3　压缩系数的确定

使用压缩自编码网络时，需要对约束项中雅克比矩阵的系数 λ 的大小进行选择，合适的比例系数可以使得样本提取的特征具有更好的局部不变性，从而提高分类的性能。图 4 给出了训练集和测试数据集为 1∶1 的情况下，参数 λ 与预测准确率之间的关系。实验结果显示，参数 λ 的取值在 0.003 附近时具有很好的分类结果，因此最终选取 λ＝0.003。

2.4　建立分类模型

实验使用 MATLABR2015b 作为编码工具，采用四层网络结构：2047-1000-200-2/4。二分类和多分类在预处理和预训练阶段完全相同，两者的差别仅仅在微调阶段，二分类的分类器选择 Logistic，而多分类的分类器选择 Softmax。详细的实验过程如下：

(1)预处理阶段：首先将所有光谱样本进行一致性处理和归一化，归一化之后数据均在[0，1]之间。

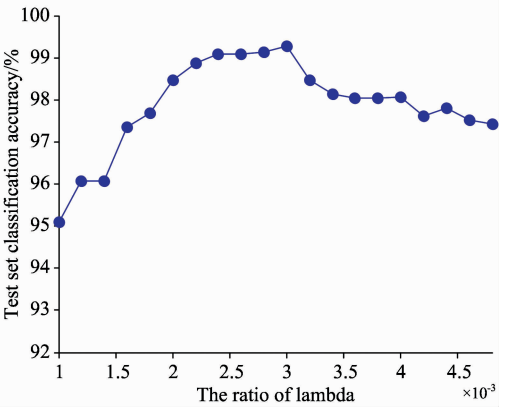

图 4　不同参数 λ 下的分类准确性

Fig. 4　The classification accuracy against different parameter lambda

(2)预训练阶段：首先训练三个 CAE，通过实验选取网络结构分别为 2074-1000-2074，1000-200-1000，200-100-200，三个 CAE 的最大迭代次数都为 200 次，学习率为 0.1，激活函数采用 sigmoid，误差函数选择均方误差。

(3)微调阶段：完成三个 CAE 的训练后，用得到的权重矩阵和偏置向量来初始化 SCAE 网络，然后通过反向传播算法对参数进行优化，微调阶段的最大迭代次数设置为 60 次，学习率为 0.05。

(4)测试阶段：用训练好的参数对测试数据集进行前馈计算，计算出每个样本属于每个类别的“概率”，所属概率最大的类别视为该样本的预测类别，然后将所有测试集中样本的预测值和真实值进行对比，得到预测准确率。

(5)对比阶段：选择 BP 神经网络、SVM 算法、SAE 和 DAE 作为对比实验。其中，BP 神经网络选择 MATLAB2015b 自带的工具箱，网络的结构与堆栈压缩自编码网络相同(2047-1000-200-2/4)。SVM 算法用 LIBSVM 工具箱，选用线性核以及高斯核函数作为对比，通过工具箱中的网格寻优法得到 SVM 的高斯核参数 C＝1，gamma＝0.32。

3　结果与讨论

3.1　二分类对比

首先，测试 SCAE 在两类药品鉴别中的预测能力。实验数据来自表 1 中的数据集 A，共收集有 252 个药品光谱样本，取广州白云制药总厂的头孢克肟胶囊共 96 个，作为实验的正类样本；取天津华津制药厂和天津医药集团津康制药厂生产的头孢克肟胶囊共 152 个，作为实验的负类样本。为了验证算法在不同大小的数据集上的预测能力，将正负样本分别按表 2 中的比例选取训练集和测试集，并且在每一个比例下分别实验 10 次得到 10 个结果，然后取 10 次结果的平均

值作为该比例下的鉴别准确率，标准差作为该比例下的算法稳定性。

分类准确性方面，在各个比例之下，BP 神经网络、线性以及高斯核的 SVM 算法，SAE，DAE，SCAE 算法在 10 次随机抽取的数据集上的平均预测准确率如表 2 所示。结果表明采用 SCAE 算法的准确率明显高于 SAE 和 DAE，说明 CAE 算法比 SAE 算法和 DAE 算法提取的特征更加具有局部不变性。从表 2 中还可见，线性核 SVM 算法的表现也很突出，特别是在训练集数量较少的情况下，说明线性核 SVM 算法适合高维数据的分类。随着训练数据集数量的增加，SCAE 算法的准确性也随之提高，并优于其他算法。

表 2 不同比例下二分类准确性

Table 2 The binary-classification accuracy on different ratios of training samples (unit: %)

Training/ Test Set	BP (2 layers)	SVM (Linear)	SVM (RBF)	SAE	DAE	SCAE
24/228	78.11	90.35	90.85	88.11	89.34	95.09
50/202	84.70	97.52	91.33	93.56	95.29	98.02
74/178	84.60	97.80	92.08	97.75	97.24	98.03
100/152	85.92	97.52	95.19	97.36	97.89	99.14
126/126	87.22	97.36	95.00	98.88	98.17	98.25
150/102	90.09	98.03	95.29	98.13	98.42	99.51
176/76	90.78	98.68	95.79	98.68	99.08	99.08
200/52	92.11	98.08	95.77	98.09	99.04	99.23
226/26	91.15	99.23	97.69	99.23	98.46	100

算法稳定性方面，基于上述实验中各个比例下测试的 10 次准确率的标准偏差(STD)如图 5 所示。SCAE 算法的稳定性与线性 SVM 算法和 DAE 算法相似，相比 BP 神经网络在稳定性上有较大的优势。随着训练数据集的增加，SCAE 的稳定性逐渐上升，且普遍优于 SAE 算法。

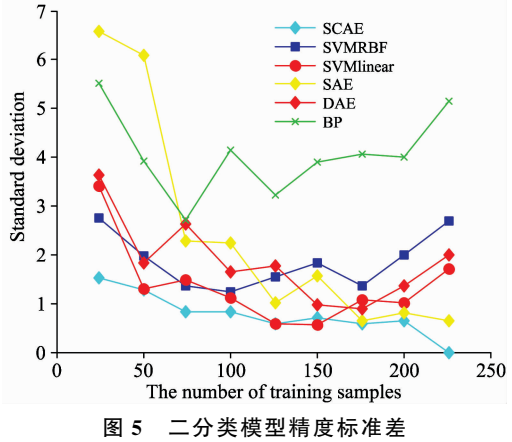

图 5 二分类模型精度标准差

Fig. 5 The standard deviation of accuracy of different binary-classification models

接下来比较运行时间，前述各算法在各比例下运行 10 次的平均时间如表 3 所示。BP 算法由于迭代次数过多，运行时间没有优势。CAE 算法加入了雅克比正则项的约束，使其往往不用达到最大迭代次数就可以找到最优解，所以在运行时间方面比 SAE 和 DAE 具有较大的优势。由于 SVM 算法不需要预训练和提取特征，所以运行时间方面比其他算法快很多。

表 3 不同比例下二分类模型运行时间

Table 3 Training time of different binary classifier on different ratios of training samples (unit: s)

Training/ Test Set	BP (2 layers)	SVM (Linear)	SVM (RBF)	SAE	DAE	SCAE
24/228	103.36	0.01	0.02	74.00	172.54	65.33
50/202	126.90	0.01	0.03	73.00	173.37	66.99
74/178	106.31	0.01	0.04	74.52	170.12	72.41
100/152	106.00	0.01	0.04	93.37	122.26	70.50
126/126	106.34	0.01	0.04	117.01	136.86	66.76
150/102	108.67	0.01	0.05	95.87	140.99	70.14
176/76	111.74	0.02	0.05	85.32	129.45	74.21
200/52	113.03	0.02	0.06	89.33	126.98	70.01
226/26	115.64	0.02	0.06	111.27	98.12	77.96

3.2 多分类对比

为了测试 SCAE 算法在多类药品鉴别的性能，实验选取四个不同厂家生产的硝酸异山梨酯片来进行鉴别。数据来源见表 1 数据集 B，共计 314 条。第一类：南京白敬宇制药有限责任公司生产，共 76 条；第二类：山西云鹏制药有限公司生产，共 89 条；第三类：太原市振兴制药有限公司生产，共 74 条；第四类：天津太平洋制药有限公司生产，共 75 条。

与二分类实验相同，为了验证各算法的分类准确率、算法稳定性和运行时间，将 4 类样本按类别随机打乱后按照表 4 中的比例构建训练集和测试集，在不同的比例下进行 10 次实验，取 10 次实验的平均准确率作为该算法的鉴别准确率。实验结果如表 4。

表 4 不同比例下多分类准确性

Table 4 The multi-classification accuracy of different classifiers on different ratios of training samples (unit: %)

Training/ Test Set	BP (2 layers)	SVM (Linear)	SVM (RBF)	SAE	DAE	SCAE
29/285	75.43	94.23	84.63	88.58	89.00	95.96
61/253	84.70	96.17	84.11	93.35	89.13	96.39
92/222	84.86	96.35	88.91	95.76	94.41	97.66
124/190	85.84	97.37	92.00	97.97	96.15	98.02
156/158	88.29	98.10	93.03	98.48	97.65	98.73
187/127	88.97	98.03	93.77	98.50	98.34	98.90
218/96	90.52	98.12	94.58	98.33	98.64	100
250/64	93.43	98.28	95.93	98.12	98.75	99.09
281/33	90.00	99.09	95.15	98.78	99.09	100

结果表明在多分类方面，在数据集较小的情况下，线性核 SVM 算法和 SCAE 算法都具有较好的分类准确率。随着训练数据量的增加，SAE 算法和 DAE 算法的分类准确率都有大幅度提高。但是，SCAE 算法的分类效果还是优于其他算法。

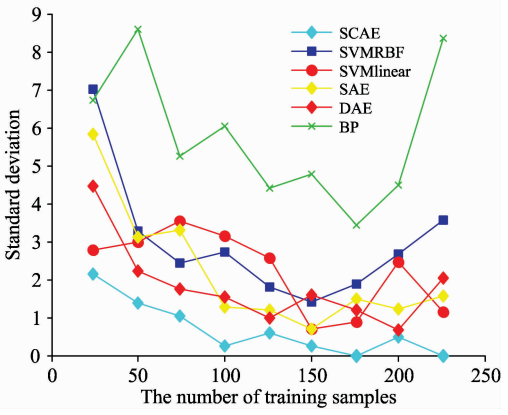

图 6　多分类模型精度标准差

Fig. 6　The standard deviation of accuracy of different multi-class classification models

表 5　不同比例下多分类模型运行时间

Table 5　Training time of different multi-class classifiers on different ratios of training samples (unit: s)

Training/ Test Set	BP (2 layers)	SVM (Linear)	SVM (RBF)	SAE	DAE	SCAE
29/285	87.33	0.01	0.04	154.96	211.87	66.52
61/253	104.79	0.02	0.07	151.87	204.76	64.67
92/222	106.58	0.02	0.10	139.79	177.31	61.69
124/190	108.72	0.03	0.13	120.52	197.45	79.72
156/158	110.83	0.03	0.17	121.88	185.80	83.16
187/127	113.20	0.03	0.21	136.33	111.45	62.54
218/96	117.35	0.04	0.25	121.84	92.35	77.17
250/64	122.33	0.04	0.31	122.16	78.73	71.29
281/33	123.03	0.05	0.35	114.37	79.51	71.86

算法稳定性方面，各算法在不同比例下 10 次预测准确率的标准差如图 6 所示。从实验结果来看，各类自编码网络普遍比 BP 算法和 SVM 算法稳定，其中 SCAE 算法相比其他的自编码网络而言更加稳定。

进一步比较运行时间，前述各算法在各比例下运行 10 次的平均时间如表 5 所示。在多分类的情况下，SCAE 算法的运行时间相比 BP，SAE 和 DAE 算法还是有一定的优势，但远慢于 SVM 算法。

4　结　论

在药品质量的监督管理中，药品鉴别是至关重要的一个环节。虽然深度网络具有很强的非线性建模能力，但是运用于近红外光谱分析时，由于光谱数据比较少，高度复杂的深度学习模型容易导致过拟合问题。本文提出的压缩自编码可以很好的提取光谱中的特征信息，学习光谱数据的内部结构特征，由于有雅克比矩阵作为约束项，使其学习到的特征更加具有局部不变性。将各层压缩自编码学习得到的参数作为堆栈压缩自编码网络各层的初始值，再通过反向传导算法对整个网络进行微调，可以有效避免整个网络陷入局部最小值，同时还可以提升网络的训练速度。从实验结果来看，使用堆栈压缩自编码对真假药品进行鉴别，其鉴别准确率普遍高于 BP 神经网络、SVM 算法、SAE 算法和 DAE 算法。在训练时间上，SCAE 快于 BP、SAE 和 DAE 算法，但远不如 SVM 算法。综合来看，运用 SCAE 算法进行药品的鉴别是有效可行的。

References

[1] LI Zhen, ZHOU Li-hong, YE Zheng-liang(李　真，周立红，叶正良). Drug Evaluation Research(药物评价研究), 2016, 39(4): 686.

[2] Fontalvo-GóMez M, Colucci J A, Velez N, et al. Applied Spectroscopy, 2013, 67(10): 1142.

[3] Deconinck E, Sacré P Y, Coomans D, et al. Journal of Pharmaceutical & Biomedical Analysis, 2012, 57(1): 68.

[4] Zou T T, Dou Y, Wang Y, et al. Science & Technology of Food Industry, 2013, 17: 317.

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

[6] Krizhevsky A, Sutskever I, Hinton G E. International Conference on Neural Information Processing Systems Curran Associates Inc., 2012. 1097.

[7] Cho K, Merrienboer B V, Gulcehre C, et al. Arxiv Preprint Arxiv, 2014, 1406: 1078.

[8] Sutskever Ilya, Vinyals O, Le Q V. Foundations & Trends® in Signal Processing, 2014, 7(3): 197.

[9] Deng L, Yu D. Foundations & Trends® in Signal Processing, 2014, 7(3): 197.

[10] Rifai S, Vincent P, Muller X, et al. Proceedings of the 28th International Conference on International Conference on Machine Learning. Omnipress, 2011. 833.

Stacked Contractive Auto-Encoders Application in Identification of Pharmaceuticals

GAN Bo-rui[1], YANG Hui-hua[1,2*], ZHANG Wei-dong[1], FENG Yan-chun[3], YIN Li-hui[3], HU Chang-qin[3]

1. College of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China
2. College of Automation, Beijing University of Posts & Telecommunications, Beijing 100876, China
3. National Institutes for Food and Drug Control, Beijing 100050, China

Abstract As near-infrared spectroscopy has many advantages, such as fast analysis, non-destructive testing and field detection, it has been widely used in many fields. However, there are some shortcomings such as low signal-to-noise ratio, weak absorption intensity and overlapping peaks in near-infrared spectroscopy. NIR spectroscopy can not be qualitatively/quantitatively obtained from the spectrum. Therefore, NIR spectroscopy can only be used as an indirect analytical technique. The research of infrared spectral modeling method becomes the core of analyzing near infrared spectroscopy. Deep learning is a new branch of machine learning and has been successfully applied in many fields. The network structure of deep learning and the non-linear activation ability make the model especially suitable for high-dimensional and nonlinear large-scale data modeling. In order to further enrich the NIRS modeling method and improve the accuracy of NIRS, it is necessary to develop a new modeling method using NIRS. The qualitative analysis of near-infrared spectroscopy is studied in this paper. A model based on Stacked Contractive Auto-Encoders(SCAE) is proposed to identify the same drugs produced by different manufacturers on the market. With contractive Auto-Encoder (CAE) based on Auto-Encoder network by adding Jacobi matrix as a constraint, self-coding network is used to reduce the dimension of the data to learn the internal characteristics of the data, and Jacobi matrix contains information in all directions. The extracted features can be invariant to a certain degree of perturbation of the input data and improve the ability of self-encoding network to extract features. SCAE is a multi-layer CAE neural network. As the input layer of the latter layer of CAE network, all the parameters of the network are obtained by adopting the layer-by-layer greedy training method. After the training, all the networks are regarded as a whole, Fine-tuning by backpropagation algorithm, and finally using Logistic/Softmax classifier for qualitative analysis. The experimental data were collected by the National Institutes for Food and Drug Control, with Cefixime Capsules as the second classification experimental data and Isosorbide Dinitrate Tablets as a multi-classification experimental data. The spectral curves were obtained by measuring the absorbance of each sample at different wavelengths with a Bruker Matix spectrometer, and then the deviation from the spectral samples was obtained by OPUS software to eliminate the drift and other factors. Next, we established the model by experimentally determining the Lamda of the constrained Jacobi matrix ratio coefficient of 0.003. The modeling process was divided into five stages, namely: pre-treatment stage, pre-training stage, fine-tuning stage, testing stage and contrast stage. In order to verify the performance of SCAE network in terms of classification accuracy, algorithm stability and modeling time, the algorithm was compared with BP neural network, SVM algorithm, sparse Auto-Encoders (SAE), Denoising Auto-Encoders(DAE) for comparison. In terms of classification accuracy, stack compression self-coding network has the highest classification accuracy and algorithm stability at different ratios of training set to test set. In terms of modeling time, SVM algorithm has a great advantage over other algorithms in terms of running time because it does not need pre-training and feature extraction. However, stack compression self-coding network modeling speed is better than other contrast algorithms except SVM. In summary, the use of stack compression self-coding network for drug identification is effective and feasible.

Keywords Stacked contractive auto-encoders; Jacobian matrix, Near infrared spectroscopy; Pharmaceutical discrimination

(Received Dec. 7, 2017; accepted Apr. 11, 2018)

* Corresponding author