

# “动态”近红外光谱结合深度学习图像识别和迁移学习的模式识别方法研究

孙禧亭, 袁洪福\*, 宋春风

(北京化工大学 材料科学与工程学院, 北京 100029)

**摘要:** 该文以山羊绒与山羊绒/羊毛混纺织物以及纯棉与丝光棉织物为研究对象, 使用其“动态”光谱, 扩大类间的光谱差异信息, 通过融合其同步和异步二维相关光谱, 用多张动态光谱构造一张能反映细节化学差异信息的“化学图像”。使用 GoogLeNet 深度神经网络图像识别模型结合迁移学习, 建立了一种光谱分类的新方法。收集了 234 个织物样品, 制备含水量分别为 0、5.4%、11.2% 和 16.3% 的样本, 同时采集样品的漫反射近红外光谱。使用干基样品的多种预处理光谱, 利用线性分类方法簇类独立软模式识别(SIMCA)和非线性方法支持向量机(SVM), 共建立了 16 个分类模型。其中, 山羊绒与山羊绒/羊毛混纺织物的 SIMCA 和 SVM 最优预测正确率分别为 63.33% 和 70.09%, 纯棉与丝光棉织物的分别为 71.02% 和 72.51%, 均不能实现有效分类。新方法对山羊绒与山羊绒/羊毛混纺织物的预测正确率为 92.59%, 纯棉与丝光棉织物的为 94.74%, 获得了有效分类。该文首次将图像分类方法用于光谱分类识别, 开辟了一种新的研究途径。针对实际应用能收集到的样品属于小样本, 不能满足深度学习需要大数据样本的问题, 使用迁移学习方法使深度学习框架适应了光谱分类(小样本), 为人工智能领域中先进的识别技术用于解决化学问题提供了一个成功示范。

**关键词:** 近红外光谱; 模式识别; 深度学习; 迁移学习

**中图分类号:** O657.33    **文献标识码:** A    **文章编号:** 1004-4957(2020)10-1247-07

## Study on Pattern Recognition Method Using ‘Dynamic’ NIR Spectroscopy with Deep Learning-based Image Identification and Transfer Learning

SUN Xi-ting, YUAN Hong-fu\*, SONG Chun-feng

(College of Material Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China)

**Abstract:** A new classification method based on chemical image was established using ‘dynamic’ near-infrared( NIR) spectroscopy with a deep learning-based image recognition model GoogLeNet and transfer learning, with cashmere, cashmere/wool blends textiles, cotton and silk cotton textiles as the targets. Moisture perturbation was proposed to apply in this paper, collecting ‘dynamic’ spectra, expanding the spectral differences between samples of different types, and thus fusing the synchronous and asynchronous two-dimensional map of dynamic spectra into a ‘chemical image’ which reflects the detailed differences between samples. A total of 234 textile samples were collected, and the samples with water contents of 0, 5.4%, 11.2% and 16.3% were prepared. Several preprocessing methods were employed before modeling. A total of 16 classification models were established, in which the best SIMCA and SVM models for cashmere vs cashmere/wool blends have the accuracies of 63.33% and 70.09%, while those of cotton and silk cotton textiles are 71.02% and 72.51%, respectively. Results demonstrated that the developed method is effective, the overall prediction correct rates of models are 92.59% for cashmere and blended and 94.74% for cotton and silk cotton. This contribution provides a successful demonstration for advanced identification techniques in the field of deep learning for solving chemical problems.

**Key words:** near-infrared spectroscopy; pattern identification; deep learning; transfer learning

收稿日期: 2020-03-17; 修回日期: 2020-06-14

基金项目: 国家重点研发计划(2016YFF0102504)

\* 通讯作者: 袁洪福, 博士, 教授, 研究方向: 光谱过程分析, E-mail: hfyuan@mail.buct.edu.cn

山羊绒与山羊绒/羊毛混纺织物, 以及纯棉与丝光棉织物均为化学组成非常接近且形态复杂的样品, 且类内变化较大, 同类织物在捻度、质地结构、染整工艺等方面均有明显的差异。由于类间化学组成变化引起的近红外光谱差别相对小, 而类内形态变化引起的光谱差异较大, 即噪声对有用信息的淹没作用较大, 导致采用“静态”近红外光谱结合常用的多元分析方法, 难以将其有效区分, 是近红外光谱分析领域中尚未解决的分类型难题<sup>[1]</sup>。

近红外光谱判别分析过程一般包括光谱预处理、特征提取和模式识别等步骤。常用的光谱预处理方法包括导数<sup>[2]</sup>、多元散射校正(MSC)<sup>[3]</sup>、标准正态变量变换(SNV)<sup>[4]</sup>、小波变化等<sup>[5]</sup>。特征提取包括主成分分析(PCA)<sup>[6]</sup>、线性判别分析(LDA)<sup>[7]</sup>、独立成分分析(ICA)<sup>[8]</sup>和基于流型学习等方法<sup>[9]</sup>。簇类独立软模式识别(SIMCA)、偏最小二乘判别分析(PLS-DA)和支持向量机(SVM)<sup>[10]</sup>是最常用的光谱模式识别方法。这些识别方法对于光谱差异较为明显的样品是有效的, 已被广泛用于诸多领域<sup>[11-12]</sup>。但对于上述化学组成差异小、形态或环境因素变化干扰大的样品, 基于近红外光谱的上述判别方法难以获得满意效果。因此, 研究解决上述问题的方法具有重要的理论和应用意义。

本文旨在建立一种全新的光谱分类方法, 采用外部扰动光谱和二维相关光谱方法, 构造能够区分样品细微差别的“化学图像”, 借助同时具有自动特征提取与分类功能的深度卷积神经网络(Convolutional neural network, CNN)方法对化学图像数据进行分类<sup>[13-14]</sup>。采用迁移学习方法, 将实际成熟的 CNN 图像分类模型适应于本研究的近红外光谱分类问题, 以期解决近红外光谱分类识别形态复杂且组成高度相似的不同类样品的分类难题。

## 1 实验部分

### 1.1 实验方法

对于化学组成高度相近且形貌差异大的不同类样品, 凭借其“静态”光谱所包含的有效分类信息量, 使用常用的多元分析判别方法不足以将它们有效分类。基于它们对外部扰动(如纤维吸水性)存在差异的现象, 对样品施加(水分)扰动, 同时采集其随外部扰动产生的“动态”光谱。一般可根据样本特性, 选择施加不同的外部扰动, 如温度、电压、磁场或化学扰动。采用它们的二维相关光谱图构造一张化学图像。与单张“静态”红外光谱(向量)相比, 新构造的化学图像明显增加了数据尺寸(矩阵)和信息量, 可放大样品的类间光谱差异。常用的光谱分类方法使用的特征提取和分类方法(适合处理向量)不再适用于化学图像(矩阵)分类。如图 1 所示, GoogLeNet 是 Google 开发的一种深度 CNN 框架<sup>[15]</sup>,

同时具有特征提取和图像识别功能, 在 ImageNet 数据集上取得了超越人眼的识别正确率, 是当今最先进和最成熟的深度学习方法之一。本文提出采用 GoogLeNet 对化学图像数据进行分类。深度 CNN 网络需使用大数据样本进行训练, 而光谱分析能收集到的样品通常属于小样本数据, 难以满足其训练要求。本文将 GoogLeNet 的图像识别模型(在 1 000 类图像大数据库中学习到的特征提取能力)向化学图像分类迁移<sup>[16]</sup>, 使用化学图像数据仅对 GoogLeNet 网络最后 4 层结构进行训练, 保留其它层结构不变, 使迁移后的 GoogLeNet 适应于当前研究的光谱分类, 从而实现形态复杂且化学组成高度相近的不同类样品的有效判别。

### 1.2 样 品

本文使用两个组分高度相近的不同类样品分类与识别案例为研究对象: ①山羊绒与山羊绒/羊毛混纺织物及纯羊毛织物的识别; ②纯棉和丝光棉织物的识别。从中国几个省份的织物生产企业和市场共收集 234 个织物实际样品。其中, 64 个山羊绒, 70 个山羊绒/羊毛混纺(羊毛含量为 51.5%~100%,

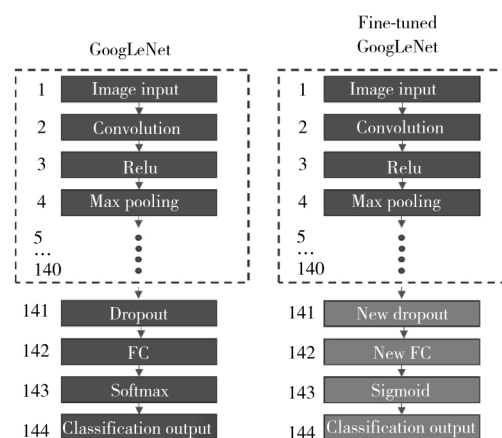


图 1 GoogLeNet 网络架构和迁移学习示意图  
Fig. 1 The architecture of GoogLeNet model and its transfer learning

包括 20 个纯羊毛织物), 50 个纯棉和 50 个丝光棉织物。它们在厚度、颜色和质地(机织或针织)上各不相同。每个样品的类别由显微镜分析法确定。其中, 山羊绒/羊毛混纺样品中的羊毛含量采用国标方法 GB/T 2910.4-2009<sup>[17]</sup> 测得。根据样品的近红外光谱, 采用 Kennard-Stone 方法<sup>[18]</sup> 对每类样品分集, 其中训练集占 80%, 剩余样品作为验证模型判别能力的测试集。

### 1.3 不同含水量样品的制备

首先将样品在 105 °C 真空烘箱中连续烘干 3 h 得到干基样品, 再将干基样品放入相对湿度(RH) 100%、恒温 20 °C 的密室中吸潮。在不同吸附时间取出, 制备不同含水量的样品。采用精度为 0.1 mg 的分析天平称重, 并计算标准条件(25 °C 和 RH 65%) 下样品的回潮率(含水量):  $\text{回潮率} = (W_e - W_d) / W_d \times 100\%$ 。其中,  $W_e$  和  $W_d$  分别为样品吸水后的重量和干重。样品的最大回潮率设定为 16.3%。每个样品制备 4 个不同的含水量(包括 0、5.4%、11.2% 和 16.3%), 共得到 936 个不同含水量的样品。

### 1.4 光谱采集

使用配有积分球附件的 Nicolet Antaris II FT-NIR 光谱仪在恒温恒湿的条件下采集样品的漫反射近红外光谱。内置的金箔用于采集背景光谱。将每块大小约 0.5 m<sup>2</sup> 的样品折叠成 4~6 层后, 直接放置在积分球的窗口上并用铁块压住, 使其与窗口表面紧密接触。光谱采集参数为: 分辨率 4 cm<sup>-1</sup>, 扫描数 32, 光谱范围 10 000~4 000 cm<sup>-1</sup>。每次光谱采集约 1 min, 每个样品重复采集 3 张光谱, 计算平均光谱作为该样品的光谱。

## 2 结果与讨论

### 2.1 光谱分析

山羊绒和羊毛纤维在组成上高度相近, 均由蛋白质和脂类等组成。图 2A 显示了所有烘干山羊绒、山羊绒/羊毛混纺织物的近红外光谱。不同样品之间的光谱基线严重漂移, 主要原因包括: 样品质地差异引起的光散射效应不同, 颜色差异引起的吸收不同。对图 2A 进行二阶导数处理, 有效消除了基线漂移, 且光谱特征峰表观分辨率获得明显改善(图 2B)。可以看出, 一方面山羊绒织物与山羊绒/羊毛混纺织物的近红外光谱含有丰富的组成信息, 另一方面, 这两种织物的近红外光谱很相近。纯棉和丝光棉织物的情况也类似, 主要组成均为天然纤维素, 其烘干样品的近红外光谱及二阶导数光谱如图 3A 和图 3B 所示。

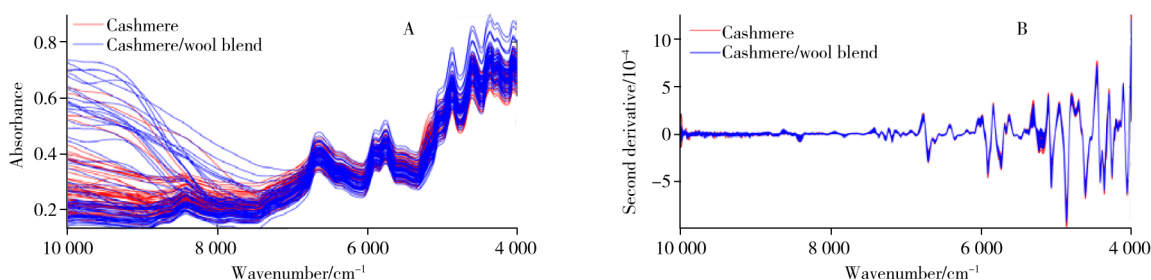


图 2 烘干山羊绒和山羊绒/羊毛混纺织物样品的原始近红外光谱(A) 及二阶导数光谱(B)

Fig. 2 Raw NIR spectra of the dried cashmere and cashmere/wool blend samples(A) and their 2<sup>nd</sup> derivative spectra(B)

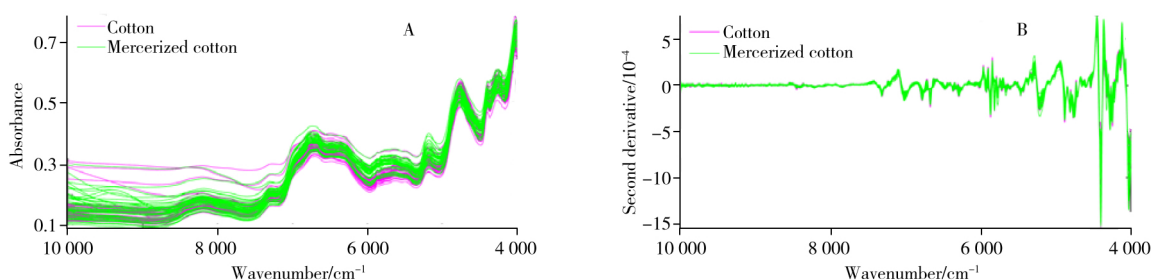


图 3 烘干纯棉和丝光棉织物样品的原始近红外光谱(A) 及二阶导数光谱(B)

Fig. 3 Raw NIR spectra of the dried cotton and mercerized cotton samples(A) and their 2<sup>nd</sup> derivative spectra(B)

## 2.2 线性方法分类

在近红外光谱定性分析常用的线性分类方法中, SIMCA 方法的效果较好, 实际应用也较多。在 SIMCA 中, 首先建立每类样本的主成分光谱空间(模型), 然后计算被测样本与各类主成分光谱空间的距离, 根据该距离确定其归属。本文建立的 SIMCA 模型中, 分别采用 MSC、一阶导数和二阶导数不同光谱预处理方法, 留一交互验证法结合  $Q$  vs Hotelling's  $T^2$  图, 确定每类模型的最优主成分数, 逐一建立了烘干山羊绒与山羊绒/羊毛混纺织物、烘干纯棉和丝光棉织物的多个 SIMCA 模型, 其统计参数如表 1 和表 2 所示。与原始光谱相比, 使用 3 种预处理方法后, 模型的预测正确率均有明显提高。其中, 二阶导数预处理方法取得了最优结果: 山羊绒与山羊绒/羊毛混纺织物的预测正确率分别为 60.48% 和 63.33%, 纯棉和丝光棉织物的预测正确率分别为 68.89% 和 71.02%。

## 2.3 非线性方法分类

SVM 是近红外光谱判别分析常用的一种分类方法, 其解决非线性问题的能力优于线性判别方法。该方法使用核函数将线性不可分的原始数据映射到更高维空间中, 使其可分。本研究使用径向基函数(RBF)作为核函数, 通过交叉验证法生成决策函数(Decision function)以抑制过拟合。应用格子搜索法(Grid search)同时优化惩罚系数(Penalty parameter)和核系数(Kernel coefficient), 它们的取值范围分别限定在如下两个列表中:  $[0.01, 0.1, 1, 10, 100, 1\ 000]$  和  $[0.001, 0.000\ 1]$ 。表 3 列出了 SVM 模型的统计数据。与表 2 对比可知, 其模型性能与 SIMCA 模型相近, 表明 SVM 同样也不能有效区分组成高度相近的不同类织物。

综上所述, 使用传统的光谱模式识别方法, 无论线性判别方法, 还是非线性判别方法, 均难以快速准确地判别组成高度相近的不同类织物。

## 2.4 化学图像构造

本文对样品施加外部水分扰动, 从而扩大了不同类样本的差异信息。以烘干样品为基础, 分别制备水含量为 5.4%、11.2%、16.3% 的样品, 并采集其水分扰动近红外光谱。从每类样品中随机选 1 个样品, 它们在不同水含量下的“动态”光谱如图 4 所示。可以看出, 与干燥样品相比, 潮湿样品的近红外光谱图中均出现 2 个较宽的水峰。其中,  $7\ 100 \sim 6\ 800\ \text{cm}^{-1}$  波段归属于  $\nu_1 + \nu_3$  振动模式( $\nu_1$  为对称伸缩振动,  $\nu_3$  属于非对称伸缩振动), 吸光度变化最大的  $5\ 150 \sim 4\ 950\ \text{cm}^{-1}$  波段归属于  $\nu_2 + \nu_3$  振动模式( $\nu_2$  为弯曲振动)。由此可见, 水分扰动显著增加了光谱数据量和信息量, 放大了不同类样品之间的近红外光谱差异。

二维相关光谱分析是一种提取扰动光谱变化信息的有力方法, 广泛用于复杂体系分析。为此, 对不同类织物的水分扰动光谱进行二维相关分析, 得到同步二维相关光谱图(图 5A1~5D1)和异步二维相关光谱图(图 5A2~5D2)。在视觉上, 与近红外光谱相比, 二维相关光谱图更加明显地反映了组成高度相近的不同织物的光谱信息差异。

表 1 山羊绒-山羊绒/羊毛混纺织物 SIMCA 模型的统计指标

Table 1 The statistics of SIMCA models for cashmere and cashmere/wool blend textiles respectively using the raw spectra and the pretreated spectra by different pretreating methods

Preprocessing	Number of latent variables		Accuracy/%	
	Cashmere	Blend	Cashmere	Blend
Raw(原始光谱)	4	7	46.19	36.67
MSC(多元散射校正)	7	5	50.71	56.67
1 <sup>st</sup> derivative	5	6	55.71	60.00
2 <sup>nd</sup> derivative	5	9	60.48	63.33

表 2 纯棉和丝光棉织物 SIMCA 模型的统计指标

Table 2 The statistics of SIMCA models for cotton and mercerized cotton textiles respectively using the raw spectra and the pretreated spectra by different pretreating methods

Preprocessing	Number of latent variables		Accuracy/%	
	Cotton	Mercerized cotton	Cotton	Mercerized cotton
Raw	4	6	51.71	52.67
MSC	8	7	55.19	56.67
1 <sup>st</sup> derivative	9	11	65.71	61.22
2 <sup>nd</sup> derivative	4	7	68.89	71.02

表 3 织物样品 SVM 模型的统计指标

Table 3 The statistics of the SVM models for textiles

Kernel coefficient	Preprocessing	Penalty parameter	Accuracy/%	
			Cashmere/blend	Cotton/mercerized cotton
0.000 1	2 <sup>nd</sup> derivative	0.1	66.67	65.71
0.001	2 <sup>nd</sup> derivative	1	70.09	72.51

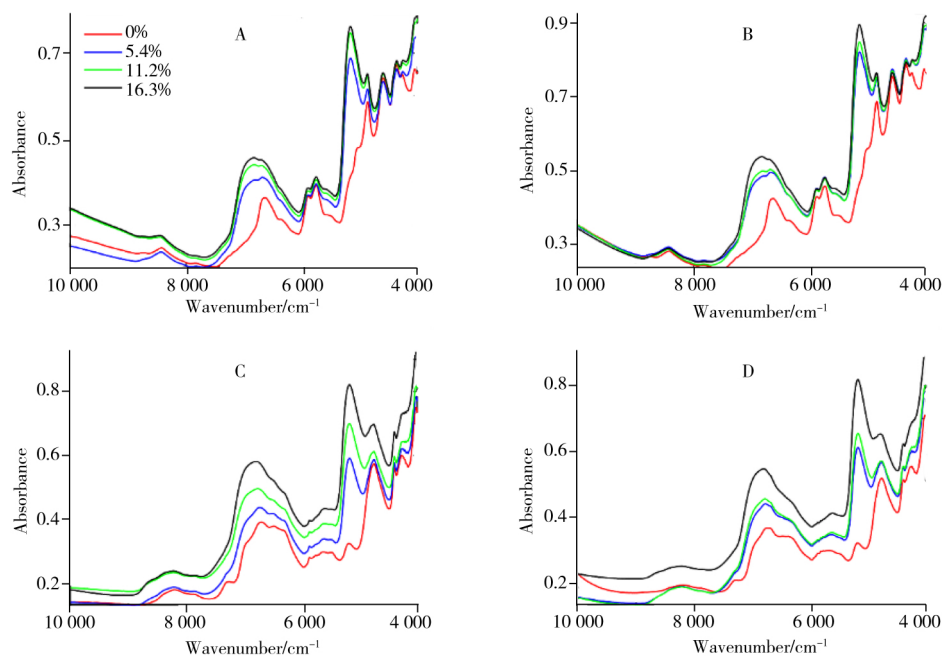


图 4 含水织物的近红外光谱图

Fig. 4 Moisture-dependent NIR spectra for a randomly selected textile

A: cashmere , B: cashmere/wool blend , C: cotton , D: mercerized cotton

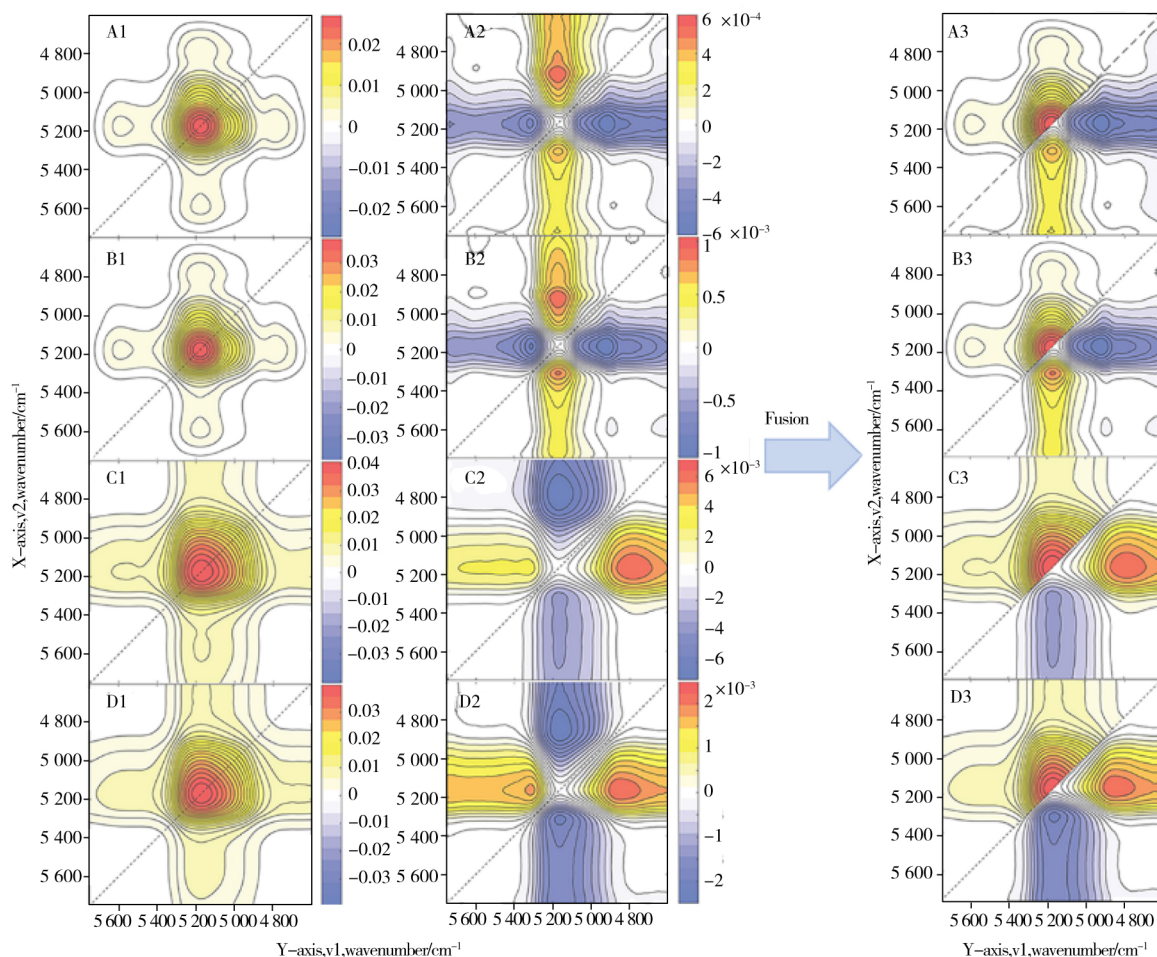


图 5 山羊绒 (A)、山羊绒/羊毛混纺 (B)、棉 (C) 和丝光棉织物 (D) 的同步相关图 (1)、异步相关图 (2) 与融合相关图 (3)  
 Fig. 5 Synchronous maps (1), asynchronous maps (2) and fusion maps (3) for cashmere (A), cashmere/wool blend (B), cotton (C) and mercerized cotton (D) textiles, respectively

由于二维相关光谱图具有对称性,为此,对同一个样品,取其同步图主对角线上半部分和异步图主对角线下半部分,合成一张既反映同步相关变化又包含异步相关性的融合光谱化学信息图像,如图 5A3~5D3 所示。与近红外光谱图相比,在视觉上,融合化学图像能更直观和更明显地反映山羊绒与山羊绒/羊毛混纺织物的光谱差异。

## 2.5 新方法分类

从信息量角度看,上述化学图像数据更有利于组成高度相似样品的分类鉴别。但常用的光谱(向量)模式识别方法,并不适合直接处理化学图像(矩阵)数据。本文提出通过迁移学习方法将 GoogLeNet 图像识别模型进行迁移,使其适合组成高度相近的不同类织物化学图像的分类与识别。

迁移学习过程需要设置各种超参数(Ultra-parameters)。其中,InitialLearnRate 指定了在损失函数的负梯度方向上的初始步长,MiniBatchSize 是在每次迭代中使用的训练集子集的大小,MaxEpochs 表示用于训练的最大 epoch 数。目前尚无调节神经网络超参数的通用规则。本研究通过 trial-and-error 方法确定了 GoogLeNet 模型的 6 个超参数(InitialLearnRate、 $1e-4$ 、MiniBatchSize、15、MaxEpochs、20)。使用训练集样品训练模型,然后通过用迁移后的 GoogLeNet 模型预测验证集样品来评估模型的性能。山羊绒与山羊绒/羊毛混纺织物以及纯棉与丝光棉织物分类模型的训练和验证过程如图 6 所示,随着 epoch 增大,训练集的交互验证正确率不断上升,10 个 epoch 后稳定在 90% 左右;而损失函数的误差值不断减小。经过 20 个 epoch 的迭代训练后,在验证集上,山羊绒织物与山羊绒/羊毛混纺织物的总体判别正确率为 92.59%,纯棉/丝光棉织物的为 94.74%(表 4),具备了实际应用的价值。因此,将在大数据库训练的 GoogLeNet 模型经迁移学习后,用于组成高度相近不同类织物的判别是可行的。

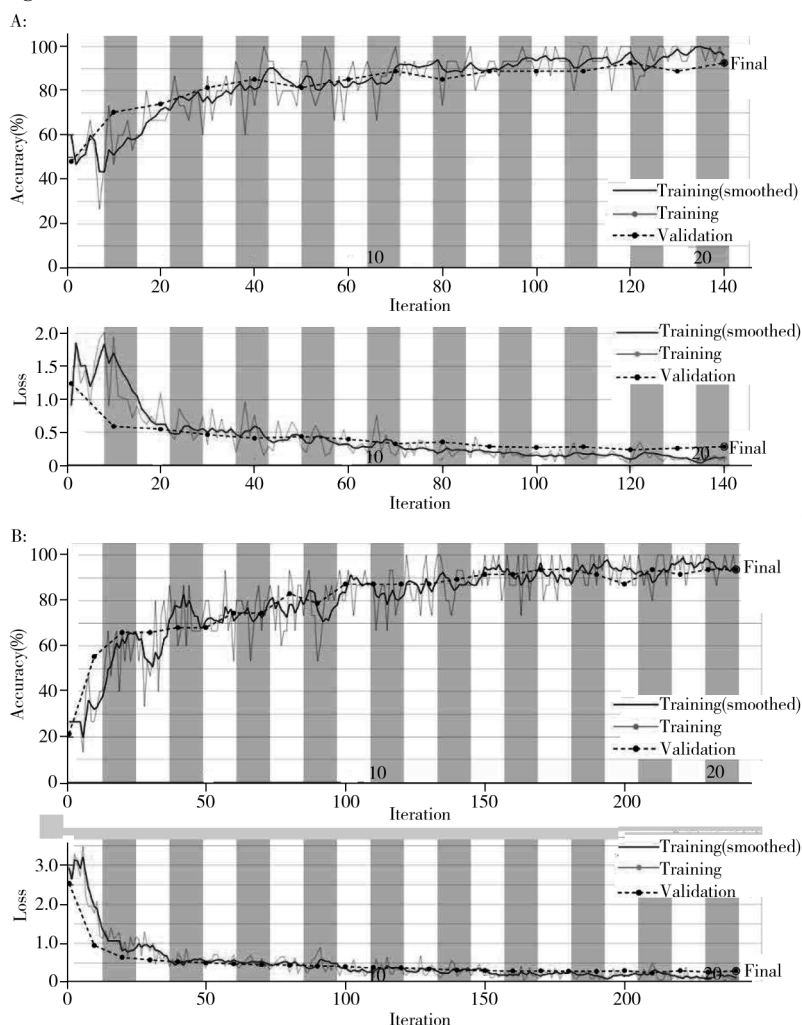


图 6 山羊绒和山羊绒/羊毛混纺(A)以及棉和丝光棉(B)迁移学习模型的训练过程

Fig. 6 Training progress and validation results for cashmere – cashmere/wool blend textile( A) and cotton – mercerized cotton textile( B) classification model



表 4 织物样品迁移学习模型的统计指标  
Table 4 The statistics of the transfer learning models for textile samples

Textile sample	Number of calibration sample	Number of corrected prediction	Accuracy/%
Cashmere	13	12	92.31
Blend	14	13	92.85
Cotton	9	9	100
Mercerized cotton	10	9	90.00

### 3 结 论

本文通过采集样品的“动态光谱”，以二维相关光谱构造化学图像，使用深度卷积神经网络结合迁移学习，建立了一种光谱分类与识别方法。研究结果表明，新方法对山羊绒与山羊绒/羊毛混纺织物的分类正确率为 92.59%，纯棉与丝光棉织物的为 94.74%，实现了对织物的高精度识别。“动态”光谱结合深度学习图像识别和迁移学习方法，解决了近红外光谱过程分析中形态复杂且组成高度相似的不同类样本的分类难题，进一步拓宽了近红外光谱分析方法的适用范围，具有重要的理论意义和实际价值。

#### 参考文献:

- [1] Zysset C, Nasser N, Büthe L, Münzenrieder N, Kinkeldei T, Petti L, Tröster G. *Opt. Express*, **2013**, 21(3): 3213–3224.
- [2] Zimmermann B, Kohler A. *Appl. Spectrosc.*, **2013**, 67(8): 892–902.
- [3] Afseth N K, Kohler A. *Chemom. Intell. Lab. Syst.*, **2012**, 117: 92–99.
- [4] Bi Y M, Yuan K L, Xiao W Q, Wu J Z, Shi C Y, Xia J, Chu G H, Zhang G X, Zhou G J. *Anal. Chim. Acta*, **2016**, 909: 30–40.
- [5] Shao X G, Leung A K M, Chau F T. *Acc. Chem. Res.*, **2003**, 36(4): 276–283.
- [6] Dai Y M, Wu Y Q, Song Y. *Infrared Phys. Technol.*, **2016**, 77: 421–430.
- [7] Esteki M, Shahsavari Z, Simal–Gandara J. *Food Control*, **2018**, 91: 100–112.
- [8] Pasadakis N, Kardamakis A A. *Anal. Chim. Acta*, **2006**, 578(2): 250–255.
- [9] Du B, Zhang L P, Zhang L F, Chen T, Wu K. *Int. J. Fuzzy Syst.*, **2012**, 14(2): 272–277.
- [10] Barker M, Rayens W. *J. Chemom.*, **2013**, 17: 166–173.
- [11] Chu X L, Yuan H F, Lu W Z. *Prog. Chem.* (褚小立, 袁洪福, 陆婉珍. 化学进展), **2004**, 16(4): 528–542.
- [12] Chu X L, Shi Y Y, Chen P, Li J Y, Xu Y P. *J. Instrum. Anal.* (褚小立, 史云颖, 陈瀑, 李敬岩, 许育鹏. 分析测试学报), **2019**, 38(5): 603–611.
- [13] Yang J, Xu J F, Zhang X L, Wu C Y, Lin T, Ying Y B. *Anal. Chim. Acta*, **2019**, 1081: 6–17.
- [14] Acquarelli J, van Laarhoven T, Gerretzen J, Tran T, Buyden L, Marchiori E. *Anal. Chim. Acta*, **2017**, 954: 22–31.
- [15] Sidorov O. *Appl. Artif. Intell.*, **2020**, 34(9): 643–655.
- [16] Weiss K, Khoshgoftaar T, Wang D D. *J. Big Data*, **2016**, 3(1): 9.
- [17] Zhu H, Shi H, Wang Z J. *Text. Auxiliaries* (朱华, 石红, 王宗吉. 印染助剂), **2011**, (7): 51–52.
- [18] Kennard R W, Stone L A. *Technometrics*, **1969**, 11: 137–48.

(责任编辑: 丁 岩)