第39卷，第7期
2019年7月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 39，No. 7，pp2142-2146
July，2019

纺织品近红外光谱定性分析的一种新方法

李海洋，刘 胜*

北京林业大学理学院，北京 100083

摘 要 近红外光谱分析技术可用于对样本的快速无损检测，在人们的生产和生活中发挥着越来越重要的作用。支持向量机是建立定性分析模型的常用方法，可通过寻找最优分类超平面将两类样本分开。在小样本情况下，支持向量机方法有其独特的优势。主成分分析是常用的数据降维方法，可将数据降维之后作为支持向量机方法的输入变量，简化模型并提高模型识别的准确性。因此，基于主成分分析的支持向量机(简称PCA-SVM)适合用于建立近红外光谱定性分析模型。多模型方法是人们使用较少的建模方法，用该方法建立的模型一般具有较好的稳定性。将多模型方法与PCA-SVM方法成功结合形成了新方法。以棉锦混合、棉涤混合纺织品为例，用新方法建立了这两类纺织品样本的近红外光谱定性分析模型。建模时将光谱数据按照波长分为4组，用每组光谱数据建立一个子模型，将子模型的输出值进行加权平均便得到最终的预测结果。这样可以更充分地使用光谱数据中所包含的信息。为了便于对比不同的方法，仍使用上述校正集和验证集，又用PCA-SVM方法建立了这两类纺织品样本的近红外光谱定性分析模型。对预测结果做交叉验证，用新方法所建模型判别的正确率的平均值为85.49%，正确率的标准差为0.066 7，用PCA-SVM方法所建模型判别的正确率的平均值为83.34%，正确率的标准差为0.109 6。研究结果表明用新方法所建模型的分类效果好于用PCA-SVM方法所建模型的分类效果；用新方法建立的模型的稳定性明显高于用PCA-SVM方法建立的模型的稳定性。用PCA-SVM方法所建模型的预测效果受校正集构成情况的影响较大，而用新方法所建模型的预测效果则相对稳定。对废旧纺织品进行分类回收可大量节约纺织原材料，但采用人工分拣方式效率低且成本高。采用近红外光谱分析方法对纺织品进行分类，为废旧纺织品的大规模精细分拣和分级奠定了一定的基础。该新方法有望用于某些其他类型样本的分类。

关键词 近红外光谱；定性分析；新方法；纺织品

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)07-2142-05

引 言

近红外光谱分析技术已被广泛应用于石化、食品、制药、农业等领域[1-4]，在人们的生产和生活中发挥着越来越重要的作用。支持向量机是近红外光谱定性分析中常用的一种方法，通过寻找最优分类超平面将两类样本分开，在小样本情况下有着独特的优势[5]。主成分分析是一种常用的数据降维方法，可将数据降维之后作为支持向量机方法的输入变量，简化模型并提高模型判别的正确率[6-9]。因此，基于主成分分析的支持向量机(PCA-SVM)方法很适合用于建立近红外光谱定性分析模型。多模型方法是目前使用较少的一种建模方法，用此方法建立的模型一般具有较好的稳定性。文献[10]利用多模型共识偏最小二乘法建立了新生儿苯丙酮尿症的红外光谱筛查模型，将模型与用偏最小二乘法建立的模型进行对比，发现前者预测更准确，稳定性也更好。有研究探讨了相思树的酸溶木质素含量预测问题，在多模型方法基础上，用预测误差较小的Klason木质素含量协助构建了酸溶木质素的近红外光谱定量分析模型，改进了酸溶木质素含量的预测效果。

本研究将多模型方法与PCA-SVM方法成功结合形成了新方法。以棉锦混合、棉涤混合两类纺织品为例，建立了上述两类纺织品样本的近红外光谱分类模型。该模型预测精度高于用PCA-SVM方法建立的近红外光谱分类模型的预测精度，且模型明显具有更好的稳定性。纺织品的生产需要消耗大量的天然纤维，如果能根据废旧纺织品所含成分对其进行

收稿日期：2018-05-25，**修订日期**：2018-09-18

基金项目：国家自然科学基金项目(61571002)资助

作者简介：李海洋，女，1993年生，北京林业大学理学院硕士研究生 e-mail：haiyanglyys@163.com

*通讯联系人 e-mail：lshlxc@163.com

分类回收，使废旧纺织品的某些成分得到重新利用，可大量节约纺织原材料。我国对废旧纺织品的回收目前基本上还是靠人工分拣，这种工作方式效率低且成本高，不利于对废旧纺织品进行大规模精细分拣和分级。本研究用近红外光谱分析方法对纺织品进行分类，为废旧纺织品的大规模精细分拣和分级奠定了一定的基础。

1　实验部分

1.1　样本的制备

为了能较好地建立纺织品的分类模型，建模时使用的纺织品样本应该具有代表性，且样本中棉含量所占的比例应该具有比较大的变化范围。在具体操作中，得到棉、锦、涤含量数据的方法有两种：一种是通过收集具有代表性的纺织品样本，用化学方法测出样本中棉、锦、涤的含量；另一种是将棉组分与锦或涤组分混合，通过调整棉与锦或涤组分的重量比例来得出一系列具有不同棉含量比例的样本。虽然用第一种方法获得的数据具有代表性，但这种方法化学测量过程复杂，工作量较大，试剂污染也比较严重。使用第二种方法虽然操作比较简单，但需考虑得到的样本是否具有代表性。

本工作选择了第二种方法。为了使棉花及纯棉布样本具有较好的代表性，从棉花的 7 个主要产地中选择了新疆、河南、湖北、河北 4 个产地，从 4 个次要产地中选择了山西，从其他产地中选择了湖南、山东、甘肃。选择产地时兼顾考虑了产地的地域分布，然后收集上述产地的年份为 2015 年或 2014 年的棉花及纯棉布。项目组还收集了来自于河北、广东、浙江、江苏的不同厂家，年份为 2015 年或 2014 年的锦纶布样、涤纶布样。锦纶、涤纶的布料与原料的区别主要是形态方面的差异，因此锦纶布样、涤纶布样与产地和生产厂家关系不大，这部分样本也具有代表性。将收集到的各种样本分别用植物粉碎机打成粉末，使之可通过 80 目筛。每次按预定数量用万分之一天平称取某种粉末，将棉和锦纶、或棉和涤纶进行混合，制备了 88 个棉锦混合样本和 160 个棉涤混合样本。样本的实际棉含量(或锦纶、涤纶含量)由称重所得数值确定。

1.2　仪器设备与光谱数据的采集

所用仪器为 UH4150 型近红外分光光度计，由日本 Hitachi 公司生产，具有双单色器棱镜-光栅光学系统，能够实现低噪声和低偏振测定。将按照不同重量比例配置好的纺织品样本放入仪器的样本池中，将分辨率设定为 5 nm，在 800～2 500 nm 谱区范围内对样本进行扫描，扫描速度为 1 200 nm·min^{-1}。在得到样本的初步近红外光谱数据之后，仪器会将本底光谱从样本的初步光谱中扣除，由此得到样本的最终近红外光谱数据。每个样本的光谱数据共包含 341 个反射率的值。

2　结果与讨论

2.1　基于主成分分析的支持向量机方法建模

为便于对结果做交叉验证并讨论模型的稳定性，将 248 个纺织品样本随机分为 A，B，C 和 D 四组，但要让每组正好包含 22 个棉锦混合样本和 40 个棉涤混合样本。将 A 组样本编号为 1，2，…，62，将 B 组样本编号为 63，64，…，124，将 C 组样本编号为 125，126，…，186，将 D 组样本编号为 187，188，…，248。每个样本的光谱数据可由一个 341 维的列向量来表示，设该向量的分量按波长从大到小的次序排列，设第 i 个样本的光谱数据为 $z_i=[z_i(1), z_i(2), \cdots, z_i(341)]^{\mathrm{T}}$，$i=1, 2, \cdots, 248$。

先用 A 组样本作为校正集建模，用 B，C 和 D 三组样本构成验证集对模型进行测试。要对校正集的光谱数据做主成分分析，需要先确定参加建模的主成分的个数，使用的主成分太少或太多都会影响模型判别的正确率。研究中尝试了校正集和验证集的多种划分方式，在建模时对每种划分方式尝试了逐个使用不同个数的主成分，结果发现：当参加建模的主成分的累积方差贡献率在 99.88%左右时，模型一般会有比较好的预测效果。此处选取校正集光谱数据的前 14 个主成分作为支持向量机方法的输入变量，其累积方差贡献率与 99.88%相差最小。对光谱数据做标准化处理[见式(1)—式(6)]，令

$$\mu_j=\frac{1}{62}\sum_{i=1}^{62}z_i(j)$$

$$\sigma_j=\sqrt{\frac{1}{62-1}\sum_{i=1}^{62}[z_i(j)-\mu_j]^2},\ j=1, 2, \cdots, 341 \quad (1)$$

令

$$\tilde{z}_i=\left(\frac{z_i(1)-\mu_1}{\sigma_1}, \frac{z_i(2)-\mu_2}{\sigma_2}, \cdots, \frac{z_i(341)-\mu_{341}}{\sigma_{341}}\right)^{\mathrm{T}} \quad (2)$$

设 x_i 是由第 i 个样本的前 14 个主成分构成的 14 维列向量，$i=1, 2, \cdots, 62$。则存在 14×341 的矩阵 $\boldsymbol{M}$，满足$(x_1, x_2, \cdots, x_{62})=\boldsymbol{M}(\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_{62})$。若第 i 个样本是棉锦混合样本，规定其标签 $y_i=1$，否则规定其标签 $y_i=-1$。设用于分类的超平面为 $w^{\mathrm{T}}x+b=0$，其中 w 和 x 都是 14 维列向量。为了求 w 和 b，需求解优化问题(3)

$$\min\frac{1}{2}w^{\mathrm{T}}w$$

$$\text{s. t.}\quad y_i(w^{\mathrm{T}}x_i+b)\geqslant1,\ i=1, 2, \cdots, 62 \quad (3)$$

其对偶问题是二次规划问题

$$\min\frac{1}{2}\sum_{i=1}^{62}\sum_{j=1}^{62}t_it_jy_iy_jx_i^{\mathrm{T}}x_j-\sum_{i=1}^{62}t_i$$

$$\text{s. t.}\quad \sum_{i=1}^{62}t_iy_i=0$$

$$t_i\geqslant0,\ i=1, 2, \cdots, 62 \quad (4)$$

求出问题(4)的解 t_1，t_2，…，t_{62}，则问题(3)的解为式(5)

$$w=\sum_{i=1}^{62}t_iy_ix_i,\ b=\frac{-1}{2}(\min_{y_j=1}w^{\mathrm{T}}x_j+\max_{y_j=-1}w^{\mathrm{T}}x_j) \quad (5)$$

设 $z=(z(1), z(2), \cdots, z(341))^{\mathrm{T}}$ 是待分类样本的光谱数据，令 $\tilde{z}$[见式(6)]

$$\tilde{z}=\left(\frac{z(1)-\mu_1}{\sigma_1}, \frac{z(2)-\mu_2}{\sigma_2}, \cdots, \frac{z(341)-\mu_{341}}{\sigma_{341}}\right)^{\mathrm{T}} \quad (6)$$

则用于分类的判决函数为 $g(z)=w^{\mathrm{T}}\boldsymbol{M}\tilde{z}+b$。用验证集的样本测试模型，若 $g(z_i)>0$，则判定第 i 个样本是棉锦混合样本，若 $g(z_i)<0$，则判定第 i 个样本是棉涤混合样本，$i=63$，

64，…，248。结果表明模型判别的正确率为 82.80%。

2.2 用新方法建模

将多模型方法与 PCA-SVM 方法相结合建立纺织品的分类模型。为便于将不同模型的分辨能力进行对比，仍用 A 组样本作为校正集，用 B，C 和 D 三组样本构成验证集。为确定起见，将子模型的个数取为 4。

因为 341÷4 的整数部分为 85，所以设 $z_i^k=[z_i(85(k-1)+1), z_i(85(k-1)+2], \cdots, z_i(85k))^{\mathrm{T}}$，$i=1, 2, \cdots, 248$，$k=1, 2, 3, 4$。用 $z_i^k(i=1, 2, \cdots, 62)$来建立纺织品的一个分类模型，称为第 k 个子模型。通过对 4 个子模型的输出结果进行加权平均，可得到最终的判决函数的取值。由于波长较小时反射率一般受噪声影响稍大，所以将 4 个子模型的权重系数依次取为 $q_1=0.35$，$q_2=0.35$，$q_3=0.2$，$q_4=0.1$。对每个固定的 k，需要对 85 维光谱数据 $z_i^k(i=1, 2, \cdots, 62)$做主成分分析。通过研究大量子模型的预测情况及用子模型输出结果的加权平均值得到的预测结果，发现当参加建立子模型的主成分的累积方差贡献率依次在 $\eta_1=99\%$，$\eta_2=99.5\%$，$\eta_3=99.5\%$，$\eta_4=99.6\%$时，将子模型的输出结果加权平均之后会有较好的预测效果。设光谱数据 $z_i^k(i=1, 2, \cdots, 62)$的前 p_k 个主成分的累积方差贡献率与 η_k 相差最小，设 x_i^k 是由 z_i^k 的前 p_k 个主成分构成的 p_k 维列向量，令

$$\tilde{z}_i^k=\left(\frac{z_i(85(k-1)+1)-\mu_{85(k-1)+1}}{\sigma_{85(k-1)+1}}, \frac{z_i(85(k-1)+2)-\mu_{85(k-1)+2}}{\sigma_{85(k-1)+2}}, \cdots, \frac{z_i(85k)-\mu_{85k}}{\sigma_{85k}}\right)^{\mathrm{T}} \tag{7}$$

从式(7)看存在 $p_k\times 85$ 的矩阵 $\boldsymbol{M}_k$，满足$(x_1^k, x_2^k, \cdots, x_{62}^k)=\boldsymbol{M}_k(\tilde{z}_1^k, \tilde{z}_2^k, \cdots, \tilde{z}_{62}^k)$。仿照 2.1 节中的方法建立第 k 个子模型，其分类超平面为 $w_k^{\mathrm{T}}x^k+b_k=0$，其中 w_k 和 x^k 都是 p_k 维列向量。令

$$\tilde{z}^k=\left(\frac{z(85(k-1)+1)-\mu_{85(k-1)+1}}{\sigma_{85(k-1)+1}}, \frac{z(85(k-1)+2)-\mu_{85(k-1)+2}}{\sigma_{85(k-1)+2}}, \cdots, \frac{z(85k)-\mu_{85k}}{\sigma_{85k}}\right)^{\mathrm{T}} \tag{8}$$

由式(8)看，子模型的判决函数为 $w_k^{\mathrm{T}}\boldsymbol{M}_k\tilde{z}^k+b_k$。为便于进行加权平均，将此判决函数写成下面的标准形式，见式(9)

$$\frac{w_k^{\mathrm{T}}\boldsymbol{M}_k\tilde{z}^k+b_k}{\|w_k\|}\ (\|w_k\| \text{ 是 } w_k \text{ 的长度}) \tag{9}$$

定义最终的判决函数为式(10)

$$g(z)=\sum_{k=1}^{4}q_k\frac{w_k^{\mathrm{T}}\boldsymbol{M}_k\tilde{z}^k+b_k}{\|w_k\|} \tag{10}$$

用验证集的样本测试模型，结果表明模型判别的正确率为 83.87%。

2.3 结果的交叉验证

先用 2.1 节的方法建模。如果用 B 组样本作为校正集，用 A，C 和 D 三组样本构成验证集，则模型判别的正确率为 93.55%。如果用 C 组作校正集，用 A，B 和 D 三组构成验证集，则模型判别的正确率为 68.28%。如果用 D 组作校正集，用 A，B 和 C 三组构成验证集，则模型判别的正确率为 88.71%。结合 2.1 节的结果，在校正集和验证集的 4 种不同构成情况下，用 PCA-SVM 方法所建模型判别的正确率的平均值为 83.34%，标准差为 0.109 6。

再用 2.2 节的方法建模。为了统一建模方法，始终将子模型的个数取为 4。如果用 B 组作校正集，用 A，C 和 D 三组构成验证集，则模型判别的正确率为 94.09%。如果用 C 组作校正集，用 A，B 和 D 三组构成验证正集，则模型判别的正确率为 77.96%。如果用 D 组作校正集，用 A，B 和 C 三组构成验证集，则模型判别的正确率为 86.02%。结合 2.2 节的结果，用新方法所建模型判别的正确率的平均值为 85.49%，标准差为 0.0667. 在校正集(和验证集)的 4 种不同构成方式下，各子模型判别的正确率如表 1 所示。

表 1 各子模型判别的正确率(%)

Table 1 The correct rate of discrimination of each sub model(%)

校正集	子模型 1	子模型 2	子模型 3	子模型 4
A 组样本	85.48	83.87	79.03	61.29
B 组样本	94.09	93.01	89.78	88.71
C 组样本	77.96	80.11	67.74	68.28
D 组样本	82.26	82.80	84.95	74.73

根据 2.3 节的结果，用新方法所建模型判别的正确率的平均值高于用 PCA-SVM 方法所建模型判别的正确率的平均值。其原因是新方法更充分地使用了光谱数据中所包含的信息，因此新方法好于 PCA-SVM 方法。

由于用新方法所建模型判别的正确率的标准差比用 PCA-SVM 方法所建模型判别的正确率的标准差小很多，所以用新方法建立的模型的稳定性明显高于用 PCA-SVM 方法建立的模型的稳定性。用 PCA-SVM 方法所建模型的预测效果受校正集构成情况的影响较大，例如用 C 组样本作校正集时，模型判别的正确率大幅低于平均值 83.34%，而用新方法所建模型的预测效果则相对稳定。

本文建模采用了校正集样本数少于验证集样本数的分组方式，这是基于以下两方面的考虑：(1) 在实际应用中，一个模型使用的次数可能远高于建模时使用的样本数，因此采用校正集样本数少于验证集样本数的分组方式更能体现不同建模方法在实际应用中的对比情况。(2) 支持向量机方法适用于小样本建模，因此本工作所用的样本分组方式具有合理性。

由表 1 的数据可以看出：用新方法所建模型判别的正确率高于大部分子模型判别的正确率，更高于子模型判别正确率的加权平均值。这说明经加权平均后，有些子模型的输出值的偏差在一定程度上被其他子模型的输出值的偏差纠正了。

3 结 论

用 PCA-SVM 方法建立了棉锦混合、棉涤混合两类纺织品样本的近红外光谱定性分析模型，又用新方法重新建立了

上述两类纺织品样本的近红外光谱分类模型，并将两种模型的分类效果进行了对比。结果表明：用新方法所建模型的分类效果好于用 PCA-SVM 方法所建模型的分类效果，且用新方法所建模型明显具有更高的稳定性。这种新方法有望用于某些其他类型样本的分类问题。

致谢：感谢张勇老师、姚胜博士的帮助！本文所用数据来源于浙江理工大学材料与纺织学院，在此致谢！

References

[1] Hu Changqin, Feng Yanchun, Yin Lihui. J. Near Infrared Spectrosc., 2015, 23(5): 271.
[2] LI Zheng-feng, XU Guang-jin, WANG Jia-jun, et al(李正风，徐广晋，王家俊，等). Chinese J. Anal. Chem.(分析化学), 2016, 44(2): 305.
[3] ZHUANG Xin-gang, WANG Li-li, WU Xue-yuan, et al(庄新港，王丽丽，吴雪原，等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2016, 35 (2): 200.
[4] Chalermpun Thamasopinkul, Pitiporn Ritthiruangdej, Sumaporn Kasemsumran, et al. J. Near Infrared Spectrosc., 2017, 25(1): 36.
[5] YANG Xiao-wei, HAO Zhi-feng(杨晓伟，郝志峰). Algorithm Design and Analysis of Support Vector Machine(支持向量机的算法设计与分析). Beijing: Science Press(北京：科学出版社), 2013. 15.
[6] Mu Weilei, Gao Jianmin, Jiang Hongquan, et al. Insight: Non-Destructive Testing and Condition Monitoring, 2013, 55(10): 535.
[7] Kuang Fangjun, Zhang Siyang, Jin Zhong, et al. Soft Computing, 2015, 19(5): 1187.
[8] Villa-Manríquez J F, Castro-Ramos J, Gutiérrez-Delgado F, et al. Journal of Biophotonics, 2017, 10(8): 1074.
[9] Saeed Bashiri, Abbas Akbarzadeh, Mansur Zarrabi, et al. Environmental Engineering & Management Journal, 2017, 16(9): 2139.
[10] WEI Wei-wei, WANG Wei-wei, SONG Xiang-gang, et al(魏伟伟，王伟伟，宋向岗，等). Journal of Analytical Science(分析科学学报), 2015, 31 (2): 257.

A New Method for Qualitative Analysis of Near Infrared Spectra of Textiles

LI Hai-yang, LIU Sheng*
College of Science, Beijing Forestry University, Beijing 100083, China

Abstract Near infrared spectral analysis technique can be used to detect samples quickly and nondestructively, which is playing an increasingly important role in people's production and life. The support vector machine is a commonly used method for building qualitative analysis models. It separates two kinds of samples by finding the optimal classification hyperplane. In the case of small samples, the support vector machine method has its unique advantages. The principal component analysis is a commonly used method to reduce the dimension of data. After the dimension is reduced by this method, the data is used as input variables of the support vector machine method. The model can be simplified and the accuracy of discriminating by the model can be improved in this way. So the support vector machine based on the principal component analysis (PCA-SVM for short) is suitable for establishing the qualitative analysis model of near infrared spectroscopy. The multi-model method is a modeling method seldom used by people. The model established by this method usually has good stability. The multi-model method is successfully combined with the PCA-SVM method to form a new method in this paper. With cotton and nylon blended, cotton and polyester blended textiles being taken as an example, a qualitative analysis model of near infrared spectra of these two types of textile samples is established by the new method. In modeling, the spectral data are divided into 4 groups according to the wavelengths. A sub model is established with each group of spectral data. The final prediction results are obtained by weighted average of the output values of the sub models. The information contained in the spectral data can be used more fully in this way. In order to facilitate the comparison of different methods, the aforementioned calibration set and validation set are used. A qualitative analysis model of near infrared spectra of these two types of textile samples is also established by using the PCA-SVM method in the paper. The cross validation of the prediction results show that the mean value of the correct rate of discrimination by the model built with the new method is 85.49%, the standard deviation of the correct rate of it is 0.066 7, and the mean value of the correct rate of discrimination by the model built with the PCA-SVM method is 83.34%, the standard deviation of the correct rate of it is 0.109 6. Since the mean value 85.49% is higher than the mean value 83.34%, the classification effect of the model built by the new method is better than that built by the PCA-SVM method. Since the standard deviation 0.066 7 is much smaller than the standard deviation 0.109 6, the stability of the model built by the new method is obviously higher than that built by the PCA-SVM method.

The prediction effect of the model built by the PCA-SVM method is greatly influenced by the composition of the calibration set. But the prediction effect of the model built by the new method is relatively stable. Sorting and recycling waste textiles can save a lot of textile raw materials. However, manual sorting is inefficient and costly. Classification of textiles by using the method of near infrared spectra analysis is proposed in this paper, which lays a certain foundation for large-scale fine sorting and grading of waste textiles. The new method put forward in the paper is also expected to be used for classification of some other types of samples.

Keywords Near infrared spectroscopy; Qualitative analysis; New method; Textiles

(Received May 25, 2018; accepted Sep. 18, 2018)

* Corresponding author