基于判别分析与 *K* 近邻算法对塑料吸管的红外光谱分析*

姜　红**，马　枭，杜　岩
（中国人民公安大学刑事科学技术学院，北京　100038）

摘要： 为建立一种塑料吸管物证的高效、准确分类方法，利用红外光谱法对来自全国的 4 个品牌共 42 个塑料吸管样本进行了检验。经过前期光谱预处理后，利用主成分分析法提取出了 25 个主成分，累积方差贡献率为 99.689%，并将其作为判别变量进行判别分析。判别结果区分效果良好但交叉验证正确率仅为 73.8%，有待进一步提高。因此将判别得分作为特征变量导入 *K* 值为 1 的 *K* 近邻算法中，构建起了分类正确率为 100%的 *K* 近邻算法模型，实现了对塑料吸管物证的准确分类。

关键词： 红外光谱法；判别分析；*K* 近邻算法；塑料吸管
中图分类号： O657.33；TQ320.7　　**文献标识码：** A　　**文章编号：** 1005-5770（2020）05-0112-05
doi： 10.3969/j.issn.1005-5770.2020.05.023　　开放科学（资源服务）标识码（OSID）：

Infrared Spectrum Analysis of Plastic Straws Based on Discriminant Analysis and *K*-nearest Neighbor Algorithm

JIANG Hong，MA Xiao，DU Yan
（School of Forensic Science，People's Public Security University of China，Beijing 100038，China）

Abstract: In order to establish an efficient and accurate classification method for the physical evidence of plastic straws，a total of 42 plastic straw samples from 4 brands across the country were tested by infrared spectroscopy. After pre-processing of the original spectrum，25 principal components were extracted by principal component analysis，and the cumulative variance contribution rate was 99.689%，which was used as a discriminant variable for discriminant analysis. The discriminating results are good，however，the accuracy rate of cross-validation is only 73.8%，which needs to be further improved. Therefore，the discriminant score is introduced into the *K*-nearest neighbor algorithm with *K* value of 1 as a feature variable，and the *K*-nearest neighbor algorithm model with classification accuracy rate of 100%is constructed to achieve accurate classification of the physical evidence of plastic straws.

Keywords: Infrared Spectroscopy；Discriminant Analysis；*K*-nearest Neighbor Algorithm；Plastic Straws

塑料作为一种高分子聚合物，凭借其价格低廉、种类繁多、适用范围广等特点被广泛应用于生产生活之中[1]，因此与案件相关的塑料物证的出现频率也逐年增加，塑料物证作为理化检验的重要研究对象，近年来得到了广泛的关注。红外光谱法作为一种快速、无损检测的光谱分析方法，在法庭科学中得到了广泛的应用[2-3]，毛志毅等[4]利用红外光谱法对塑料管材进行检验，结合一致性模型判断塑料管材中是否含有再生塑料。马枭等[5]结合红外光谱法和 X 射线荧光光谱法对塑料绳有机成分与无机填料进行了分析，并据此实现了对样本的分类。

塑料吸管作为直接与人体接触的塑料制品，其成分较为简单且又大致分为透明与不透明两类，因此多采用红外光谱反射法对其进行检验。作为塑料物证的一种，目前已有对塑料吸管的报道：杜岩等[6]利用红外光谱法根据谱图特征峰的不同对不同品牌的塑料吸管样本进行了区分。但传统分析方法往往更注重谱图的比对，增加了时间成本的同时也无法保证分类正确率，因此对于塑料吸管物证的红外光谱法研究有待进一步深入。

本实验将采集到的红外光谱数据进行预处理后通过主成分分析降维后进行判别分析，将判别得分作为特征进行 *K* 近邻算法分类模型的构建，达到建立一种塑料吸管物证高效、准确分类方法的目的。

1　实验部分

1.1　实验仪器

傅里叶变换红外光谱仪：Nicolet-6700，美国赛

*　中国人民公安大学 2019 年度基本科研业务费重点项目（2019JKF222），国家重点研发计划项目（2017YFC0822004）
**　通信作者：姜红，女，1963 年生，硕士，教授，主要从事微量物证分析方面的研究。jiangh2001@163.com

默飞世尔。

1.2　实验样本

收集来自全国各地 4 个品牌的塑料吸管样本共 42 个（样品表略）。

1.3　实验操作

采用 Smart Performer 采样器，扫描波数范围为 675～4 000 cm^{-1}，分辨率为 8 cm^{-1}，扫描次数为 64 次，增益为 1，Y 轴最终格式为吸光度（A）。

经前期预实验，由于有色不透明样本透射效果不佳，所测得数据信噪比较低，因此不宜采用透射法，而是利用反射法对样本进行检测。利用无水乙醇对样本进行擦拭，自然风干后放置于采样器中，调节采集头位置与松紧程度后开始测量，测量完毕后取出样本。重复上述操作完成对所有样本的红外光谱测量，得到原始红外光谱数据。

1.4　数据预处理

由于采集所得数据存在量级不同、信噪比较低等问题，为了提高数据的准确度，对原始红外数据进行 Z 分数归一化，将数据标准为 N（0，1），达到消除量级不同造成干扰的目的。利用多项式阶为 2，窗口点数为 25 的 Savitzky-Golay 平滑法对数据进行平滑处理，消除部分噪音干扰，最后进行多元散射校正处理得到谱图见图 1。

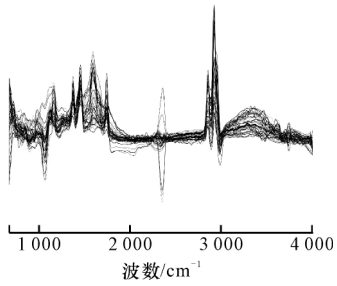

图 1　预处理后 42 个塑料吸管样本红外光谱图

Fig 1　Infrared spectrogram of 42 plastic straw samples after pre-processing

2　结果与讨论

2.1　主成分分析

主成分分析（PCA）是将高维空间中的多个指标简化为低维空间中的较少指标的一种降维方法，经过降维处理后的较少的综合指标能够解释并包含全部指标信息[7-10]。本实验中经预处理后的红外光谱数据能很好地反映出各样本之间的差异，但在实际应用中，由于每一光谱波数对应着特定的值，这种一一对应的关系导致变量数据极为庞大，并且存在着价值较低的数据特征[11]。为了达到降低分析数据复杂程度的目的，本实验利用 SPSS 25.0 分析软件对红外光谱数据进行降维处理，提取主成分总方差解释见表 1。

表 1　主成分总方差解释[1)]

Tab 1　Total variance explanation of the principal component

成分	特征值	方差贡献率/%	累积方差贡献率/%
1	516.83	29.912	29.912
2	368.627	21.357	51.269
3	307.734	17.829	69.099
4	151.703	8.789	77.888
5	101.834	5.900	83.788
6	83.869	4.859	88.647

注：1）其余主成分忽略。

为了保证所提取主成分能对全部变量准确地进行解释，主成分分析通常提取特征值大于 1 且累积方差贡献率大于 85% 的主成分[12]。由于前 6 个主成分特征值均大于 1 且累积方差贡献率为 88.647%，大于 85%，因此表 1 提取出了前 6 个主成分，其余主成分对变量解释能力较弱因而暂时忽略。

2.2　判别分析

判别分析作为一种有监督的分析方法，其中心思想为将较多维数的变量数据投影到较低维度的空间上，使得类与类之间最大程度区分开来，类内样本聚集。然后依据类间距离最大和类内距离最小的原则计算判别函数，依据计算所得的判别函数可达到对新样本进行判别分类的目的[13-14]。

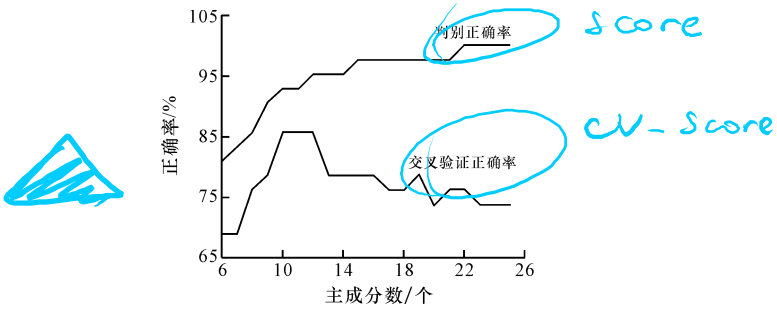

图 2　主成分判别正确率

Fig 2　The discrimination accuracy rate of principal component

本实验前期已通过主成分分析对红外光谱数据进行了降维、压缩处理，结合判别分析可进一步实现模式识别函数的构建。将所提取的累积方差贡献率为 88.647% 的 6 个主成分作为变量进行判别分析，判别正确率为 81.0%，交叉验证正确率仅为 69.0%。但随着主成分的增加，当主成分个数增加至 25 个，累积方差贡献率增加至 99.689% 时，判别正确率达到 100%，交叉验证正确率达到 73.8%，判别分析正确

率显著提升（见图 2）。

究其原因，主要是由于初始提取的 6 个主成分对原始变量的解释与包含并不全面，从而导致判别与交叉验证的正确率较低，随着主成分数量增加到 25 个时，对于原始变量的解释概括能力显著提高，判别与交叉验证的正确率显著提升[15]。因此应选用前 25 个主成分作为判别分析的变量构建判别函数，典则判别函数摘要见表 2。

表 2 典则判别函数摘要

Tab 2 Summary of canonical discriminant

函数	特征值	方差百分比/%	累积百分比/%
1	18.059	50.1	50.1
2	10.700	29.6	79.7
3	7.308	20.3	100

判别分析一共计算出 3 组判别函数，判别函数 1 方差为 50.1%，判别函数 2 方差为 29.6%，前两组判别函数累计方差已达到 79.7%，已经能够较好地对各类样本进行区分。但考虑到判别函数 3 方差为 20.3%仍占较大比重，对样本的区分应有较大的作用，因而不能将其忽略。分别将判别函数 1、2、3 作为 x、y、z 轴绘制三维判别函数得分散点图见图 3。

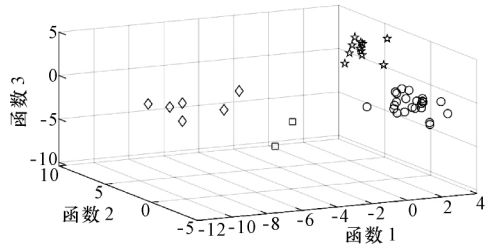

图 3 三维判别函数得分散点图

Fig 3 Three-dimensional scatter plot of discriminant function score

由图 3 可知，4 类塑料吸管样本在 3 组判别函数上明显区分开来，判别分类效果良好。虽然判别正确率达到 100%，但交叉验证正确率仅为 73.8%，可能由于选择了较多的主成分，导致信噪比降低，对交叉验证结果产生较大的影响[16]。因此，在现有基础上应探索更加合理准确的分类方法。

2.3 K 近邻分析

K 近邻算法中心思想是将各个样本之间的距离（例如欧式距离、马氏距离等）计算出来，凭借距离的大小判断出该样本与在空间内距离最近的 K 个样本之间的亲疏关系，通过最近邻的 K 个样本的类别归属判断该样本的类别[17-18]。

由于前期构建的判别模型判别正确率较高，因此使用 3 组判别函数上各主成分变量的判别得分作为 K 近邻算法的特征变量，构建 K 近邻算法模型对各样本进行分类，预测变量重要性见图 4。

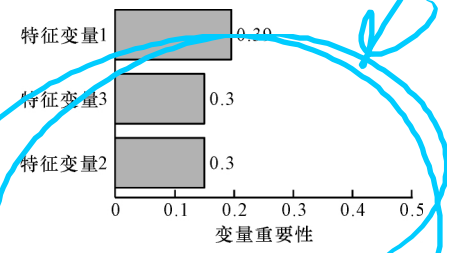

图 4 预测变量重要性

Fig 4 The importance of predictive variables

预测变量重要性越大即代表其对样本分类的贡献度越高，所有预测变量的重要性之和为 1。其中特征 1 的重要性为 0.39（保留两位小数），特征 3 和 2 的重要性为 0.3，表示在分类时特征 1 所做出的贡献大于特征 3 和 2。因此在进行 K 值选择时按照重要性对特征进行加权，重要性较大的特征占较大权重，反之则占较小权重。K 值选择结果进行交叉验证后错误率见表 3。

表 3 K 选择错误率

Tab 3 Error rate of K selection

K 值	选择错误率/%
1	0
2~8	5.33
9	7.83
10	11.17

K 值的选取是 K 近邻算法中的一个重要环节，通常 K 值最小值从 1 开始选取，最大值则应不大于样本数的平方根。本实验中共有塑料吸管样本 42 个，因此 K 值的选择范围应为 1~6，当 K 值为 1 时交叉验证后的错误率为 0，即表示未知某一样本类别时，选择距离未知样本最近的一个已知类别的样本，将其类别作为未知样本的类别，该分类方法经过交叉验证后错误率为 0。例如，35#样本在 K 值为 1 时，观测到其最近邻元素为第Ⅰ类 38#样本，最近欧氏距离为 0.048，因此将 35#样本归为第Ⅰ类，经核对，35#样本属于第Ⅰ类样本，该方法验证了 K 值为 1 时模型的准确度。但当 K 值为 2~6 时即将距离样本最近的 2~6 个样本作为最近邻元素时，K 选择错误率增加至 5.33%，表示分类结果出现一定的误差。

究其原因，前期判别分析已将 4 类样本明显区分开来，判别后类间距离显著增加，类内距离明显收缩，所以当 K 为 1 时与之相近邻的元素为同一类样本。但由于各类样本数目不均，当 K 值取值为 2~6

时，可能存在类内样本数目小于 K 值的情况，导致最近邻元素存在已明显划分为其他类别的样本，因此分类结果出现错误。

综上所述，最终选择 K 值为 1，所有样本作为训练样本进行 K 近邻分类，分类结果及正确率见表 4。

表 4　K 近邻分类正确率

Tab 4　Accuracy rate of K-nearest neighbor classification

观测值	预测值				
	1	2	3	4	正确率/%
1	23	0	0	0	100
2	0	6	0	0	100
3	0	0	11	0	100
4	0	0	0	2	100
总体百分比/%	54.8	14.3	26.2	4.8	100

由表 4 可知，42 个塑料吸管样本经 K 近邻算法分类后正确率为 100%，且 $K=1$ 经交叉验证后正确率也为 100%，表明该分类方法准确可行。在构建起已知类别样本的分类模型的基础上，可实现对于未知样本的准确分类。

3　结论

本实验利用红外光谱法对 42 个不同品牌的塑料吸管样本进行了检验，经过预处理后对数据进行降维处理，提取出 25 个特征值均大于 1，累积方差贡献率为 99.689%的主成分。利用主成分作为变量进行判别分析，但判别结果交叉验证正确率仅为 73.8%。因此提取判别得分作为 K 近邻算法特征变量，选择 K 值为 1，最终分类正确率为 100%，实现了对塑料吸管样本的准确分类。但值得注意的是在本实验中由于样本类内数量的影响，对 K 值的选择产生了一定的影响，因此在接下来的工作中，将继续探求一种样本数对分类效果影响较小或无影响的有监督算法模式，并将其应用于塑料物证的模式识别中。

参　考　文　献

[1] 李文环，金尚忠，陈玲玲，等. 基于近红外光谱结合主成分分析和 BP 神经网络的常用塑料快速鉴别 [J]. 塑料工业，2016，44 (12)：124-127，137.
LI W H，JIN S Z，CHEN L L，et al. Rapid identification of common plastics based on near-infrared spectrum with the combination of principal component analysis and BP neural network [J]. China Plast Ind，2016，44 (12)：124-127，137.

[2] 姜红，王洪波. 傅立叶变换红外光谱法检验口红的研究 [J]. 刑事技术，2010 (4)：20-24.
JIANG H，WANG H B. Analysis of lipsticks by fourier transform infrared spectrophotometer [J]. Forensic Sci Technol，2010 (4)：20-24.

[3] 姜红，韩莹. 傅里叶变换红外光谱法检验护肤品样品 [J]. 中国人民公安大学学报 (自然科学版)，2010，16 (3)：1-4.
JIANG H，HAN Y. Analysisof skin care samples by transforming infrared spectrophotometer (FTIR) with fourier [J]. J People's Public Security Univ China (Sci Technol)，2010，16 (3)：1-4.

[4] 毛志毅，滕藤，徐一飞，等. 红外光谱法鉴别塑料管材中的再生塑料 [J]. 理化检验 (化学分册)，2017，53 (12)：1370-1374.
MAO Z Y，TENG T，XU Y F，et al. Differentiation of regenerated plastics in plastic pipe materials by infrared spectroscopy [J]. Phys Test Chem Anal (Part B：Chem Anal)，2017，53 (12)：1370-1374.

[5] 马枭，姜红，杨佳琦. 红外光谱结合 X 射线荧光光谱检验塑料打包带 (绳) 的研究 [J]. 化学研究与应用，2019，31 (9)：1643-1648.
MA X，JIANG H，YANG J Q. Research on inspecting the plastic pack belts (ropes) by infrared spectrometrycombined with X-ray fluorescence spectrometry [J]. Chem Res Appl，2019，31 (9)：1643-1648.

[6] 杜岩，姜红，李晓白，等. 傅里叶变换红外光谱法检验塑料吸管的研究 [J]. 上海塑料，2014 (2)：38-42.
DU Y，JIANG H，LI X B，et al. Analysis of plastic straws by fourier transform infrared spectroscopy [J]. Shanghai Plast，2014 (2)：38-42.

[7] 林海明，张文霖. 主成分分析与因子分析的异同和 SPSS 软件——兼与刘玉玫、卢纹岱等同志商榷 [J]. 统计研究，2005 (3)：65-69.
LIN H M，ZHANG W L. The Relationship between principal component analysis and factor analysis and SPSS software—to discuss with comrade Liuyumei，Luwendai etc [J]. Stat Res，2005 (3)：65-69.

[8] RUIZ J R R，CANALS T，GOMEZ R C. Comparative study of multivariate methods to identify paper finishes using infrared spectroscopy [J]. IEEE Trans Instrument Meas，2012，61 (4)：1029-1036.

[9] 公丽艳，孟宪军，刘乃侨，等. 基于主成分与聚类分析的苹果加工品质评价 [J]. 农业工程学报，2014，30 (13)：276-285.
GONG L Y，MENG X J，LIU N Q，et al. Evaluation of apple quality based on principal component and hierarchical cluster analysis [J]. Trans Chin Soc Agric Eng，2014，30 (13)：276-285.

[10] 高惠璇. 应用多元统计分析 [M]. 北京：北京大学出版社，2005：265-272.
GAO H X. Apply multivariate statistical analysis [M].

Beijing: Peking University Press, 2005: 265-272.

[11] STANIMIROVAI, WALCZAK B, MASSART D L, et al. A comparison between two robust PCA algorithms [J]. Chemom Intell Lab Syst, 2004, 71 (1): 83-95.

[12] 许新征，丁世飞，杨胜强，等. 煤与瓦斯突出的 PCA-BP 神经网络预测模型研究 [J]. 计算机工程与应用，2011，47 (28)：219-222.
XU X Z, DING S F, YANG S Q, et al. Model for predicting coal and gas outburst based on PCA and BP neural network [J]. Comput Eng Appl, 2011, 47 (28): 219-222.

[13] 李进前，王起才，张戎令，等. 基于 Fisher 分析的高速铁路地基膨胀土判别方法 [J]. 铁道建筑，2017，57 (8)：73-77.
LI J Q, WANG Q C, ZHANG R L, et al. Discriminant method of expansive soil in high speed railway foundation based on fisher analysis [J]. Railway Eng, 2017, 57 (8): 73-77.

[14] 邵良杉，徐波. 基于因子分析与 Fisher 判别分析法的隧洞围岩分类研究 [J]. 公路交通科技，2015，32 (7)：98-104，119.
SHAO L S, XU B. Research on classification of tunnel surrounding rock based on factor analysis and fisher discriminant analysis [J]. J Highway Transportation Res Develop, 2015, 32 (7): 98-104, 119.

[15] 马枭，姜红，杨佳琦. X 射线荧光光谱结合多元统计分析塑料打包带（绳） [J]. 激光与光电子学进展，2019，56 (22)：243-247.
MA X, JIANG H, YANG J Q. Examination of plastic pack belts (ropes) by X-ray fluorescence spectra combined with multivariate statistical analysis [J]. Laser Optoelectronics Prog, 2019, 56 (22): 243-247.

[16] 何欣龙，陈利波，王继芬，等. 基于 K 近邻算法的塑钢窗拉曼光谱分析 [J]. 激光与光电子学进展，2018，55 (5)：409-413.
HE X L, CHEN L B, WANG J F, et al. Ramanspectroscopy analysis of plastic steel window based on K nearest neighbors algorithm [J]. Laser Optoelectronics Prog, 2018, 55 (5): 409-413.

[17] 张瑜，谈黎虹，何勇. 近红外透射光谱结合判别分析方法在汽车制动液品牌与新旧鉴别中的应用研究 [J]. 光谱学与光谱分析，2016，36 (10)：3179-3184.
ZHANG Y, TAN L H, HE Y. Identification of brake fluid brands, new and used brake fluid with discriminant analysis based on near-infrared transmittance spectroscopy [J]. Spectroscopy Spectral Anal, 2016, 36 (10): 3179-3184.

[18] BALABIN R M, SAFIEVA R Z, LOMAKINA E I. Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines [J]. Microchem J, 2011, 98 (1): 121-128.

（本文于 2020-02-11 收到）

（上接第 106 页）

WANG S R, LIU Q, ZHENG Z, et al. Study on the mechanism of biomass thermal cracking based on thermogravimetric analysis [J]. J Eng Thermophys, 2006 (2): 351-353.

[9] 罗希韬，王志奇，武景丽，等. 基于热重红外联用分析的 PE、PS、PVC 热解机理研究 [J]. 燃料化学学报，2012，40 (9)：1147-1152.
LUO X T, WANG Z Q, WU J L, et al. Study on the pyrolysis mechanism of PE, PS, PVC based on thermogravimetric analysis [J]. J Fuel Chem Technol, 2012, 40 (9): 1147-1152.

[10] 孙宁，陈婵，冯春云，等. 热重-傅立叶变换红外光谱联用法研究超支化聚氨酯/丙烯酸酯的热性能及热分解机理 [J]. 分析测试学报，2015，34 (8)：887-893.
SUN N, CHEN C, FENG C Y, et al. Thermogravimetric-fourier transform infrared spectroscopy study on thermal properties and thermal decomposition mechanism of hyperbranched polyurethane/acrylic acid [J]. J Instrumental Anal, 2015, 34 (8): 887-893.

[11] QB/T 2957—2008. 淀粉基塑料中淀粉含量的测定热重法（TG） [S]. 北京：中国轻工业出版社，2008.
QB/T 2957-2008. Determination of starch content in starch-based plastics Thermogravimetry (TG) [S]. Beijing: China Light Industry Press, 2008.

[12] 谢启桃，胡克良，潘忠孝，等. 小波变换用于不同晶型聚丙烯树脂红外特征峰的鉴别 [J]. 化学通报，1998 (2)：62-65.
XIE Q T, HU K L, PAN Z X, et al. Wavelet transform used to identify infrared characteristic peaks of different crystalline polypropylene resins [J]. Chemistry, 1998 (2): 62-65.

[13] 任静，刘刚，欧全宏，等. FTIR 结合 DWT 鉴别研究六种不同植物来源的淀粉 [J]. 湖北农业科学，2016，55 (5)：1277-1280.
REN J, LIU G, OU Q H, et al. FTIR combined with DWT for the identification of starch from six different plant sources [J]. Hubei Agric Sci, 2016, 55 (5): 1277-1280.

（本文于 2020-03-10 收到）