第31卷,第10期
2011年10月

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

Vol. 31, No. 10, pp2747-2752
October, 2011

一种基于拉曼光谱的石油产品快速分类方法

李 晟, 戴连奎*

浙江大学工业控制技术国家重点实验室, 浙江 杭州 310027

摘 要 提出了一种基于拉曼光谱的石油产品快速分类方法。首先, 利用经过谱图预处理的石油产品训练样本拉曼谱图构建模型知识库, 计算各类别的特征拉曼谱图和类内阈值; 其次, 将石油产品测试样本的拉曼谱图经过相同的预处理, 再计算其与各类别特征拉曼谱图的线性相关系数, 若最大相关系数大于或等于最大相关系数对应类别的类内阈值, 则该样本属于此类别。针对7类96个取自不同炼厂不同批次的石油产品样本和4个未知类别样本的分类测试表明: 该方法可正确地对常用的石油产品样本进行分类, 也可判断未知样本的存在。该方法概念简单清晰, 无需人为干涉, 不存在复杂的数学运算, 便于实际应用中的程序实现。

关键词 拉曼光谱; 石油产品; 快速分类; 相关分析
中图分类号: O657.3 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2011)10-2747-06

引 言

石油产品由原油经过不同的处理流程和各种物理化学反应得到, 种类繁多, 化学组成也各不相同。常见的石油产品有: 汽油、柴油、煤油、轻质石脑油和润滑油等等。石油产品的快速分类有着十分重要的意义。例如, 在油品长距离传输时, 为节约建设成本, 往往在一条管线上传输各种不同的石油产品, 实时地监控管线中石油产品种类以避免错误的传输和不必要的过量传输非常重要。

在传统的石油产品分类的方法中, 最直接的方法是根据石油产品的物理性质状态, 如气味、颜色、密度、固液状态等等区分, 这类方法不可避免地会产生分类错误和不可分的情况。超声波、电导率测量技术也被用来区分石油产品, 但是具有相同超声波传导速度或电导率的石油产品仍然不一定是同种石油产品。上述方法的缺陷在于它们都不能从本质上反映不同石油产品的化学组成。

光谱分析技术可以在谱图上直接体现物质化学组成信息。自上20世纪90年代以来, 在激光、光纤、微电子、计算机技术和化学计量学发展的推动下, 光谱分析技术广泛应用于农业、纺织、制药和石油化工等领域[1, 2]。

在众多的光谱分析技术中, 近红外光谱的应用最为广泛, 近红外光谱技术结合一定的化学计量学方法即可实现对石油产品的分类或指标的预测。Kim等[3]利用principal component analysis(PCA)和贝叶斯分类原理建立了基于近红外光谱的石油产品分类器; Balabin等[4, 5]将不同产地不同种类的汽油样本的近红外谱图作为数据源, 在特征提取后使用多种分类方法对这些汽油样本进行分类, 并对分类结果进行比较分析。国内, 张其可等[6]针对不同产地批次牌号的汽油样本近红外谱图, 提出了合理的定性和定量分析方法。然而, 上述的分类均存在一些分类错误。从机理角度分析, 这是由于光在不同油品的近红外波段的选择性吸收的确存在差别, 但是幅度不大, 容易受到噪声或其他未知因素的干扰。

拉曼光谱法是另一种有效的检测手段, 其对基团和化学键的体现较近红外光谱更加丰富, 差异度更大, 适合用于石油产品这类有多种烃类组成的复杂混合物的分类。Cooper等[7, 8]已经利用拉曼光谱技术进行了石油产品某些指标的定量分析。我们也将45个汽油拉曼谱图样本, 利用PCA类中心最小距离法, 实现了无错分的牌号分类。可以说, 利用拉曼光谱技术对石油产品种类进行分类也是可行的。

本研究针对石油产品的分类问题, 提出了一种基于拉曼光谱的快速分类方法。该方法将经过谱图预处理的各类石油产品拉曼谱图与模型库中各类的特征谱图进行线性相关分析, 根据最大相关系数是否超过对应类别的类内阈值决定石油产品类别, 同时也可对未知石油产品样本给出合理的判断。该方法简单, 无需复杂运算, 大大节约了模型计算的时

收稿日期: 2011-01-16, 修订日期: 2011-05-20
基金项目: 国家(863计划)项目(2009AA04Z123)资助
作者简介: 李 晟, 1985年生, 浙江大学控制系博士研究生 e-mail: sli@iipc.zju.edu.cn
* 通讯联系人 e-mail: lkdai@iipc.zju.edu.cn

间。同时，本方法已经使用 C+ + 语言实现，并成功应用于我们自主研制的石油产品快速分类仪中，取得良好的使用效果。

1 算法理论

基于拉曼光谱石油产品快速分类方法具体流程如图 1 所示，主要包括：拉曼光谱获取、光谱预处理、线性相关运算与类别判别等步骤。Y_s 为石油产品训练样本的已知类别，Y_t 为测试样本的判断类别，X_s 为经过谱图预处理的训练样本拉曼谱图，X_t 为经过谱图预处理的测试样本拉曼谱图，X_m 为模型知识库中类别特征拉曼谱图，R_{min} 为类内阈值，R 为 X_t 与所有 X_m 的相关系数中的最大值。

1.1 谱图预处理

为去除拉曼光谱谱图中与待测样本属性无关的干扰，如仪器的随机噪声、谱图中荧光背景等等，必须进行拉曼光谱的预处理。预处理的步骤主要包括：干涉光的去除、荧光背景的扣减、整数插值、平滑滤波和归一化。

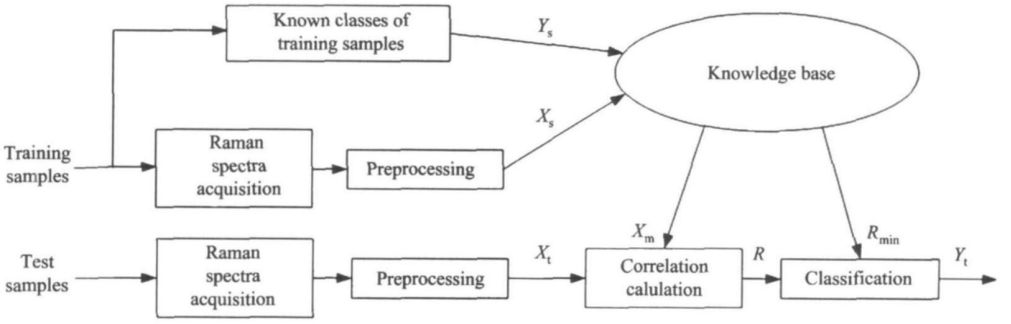

Fig 1 Data processing flow of fast classification of petroleum samples

1.2 模型知识库的构建

模型知识库是存储石油产品类别知识的数据库，该数据库存储的数据包括：经过谱图预处理的各类石油产品训练样本拉曼谱图、与拉曼谱图对应的石油产品类别、各类别的特征拉曼谱图和类内阈值。某类别的特征拉曼谱图为该类别内所有谱图的平均谱图。类内阈值计算方法如下：该类别内所有样本拉曼谱图两两进行线性相关分析，所得的线性相关系数中的最小值即为类内阈值。

1.3 线性相关分析和类别判断

对于同类的石油产品，其所含的化学成分是相对固定的，所不同的是各种成分在比例上的微小差别。这种差别在拉曼谱图的体现在于：同类石油产品的拉曼谱图具有类似的大致形状，具体差异在于一些拉曼特征峰强度的有限变化。基于这个事实，若某两个石油产品样本的拉曼谱图线性相关系数很高(接近于 1)，则证明这两个石油产品样本拉曼谱图形状十分类似，由此，可判断这两个石油产品样本属于同一类。当相关系数低于某个阈值时，则证明这两个石油产品样本属于不同类。

向量 X 和 Y 的线性相关系数的计算方法如下

$$R_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\ \sqrt{E(Y^2) - E(Y)^2}}$$

其中 $E(X)$ 代表向量 X 的数学期望，具体计算时可用向量 X 的平均值代替。

测试的石油产品样本谱图与所有类别特征谱图进行线性相关分析，选取最大线性相关系数对应类别的类内阈值与最大线性相关系数比较，若最大线性相关系数大于等于该类内阈值，则测试样本属于此类别。反之，测试样本属于未知类别。

2 实验部分

2.1 样本来源

本次实验使用的石油产品样本取自不同炼厂不同批次共 96 个样本，可分为四大类，即：柴油样本 20 个、汽油样本 48 个、轻质石脑油样本 10 个、甲醇汽油样本 18 个。根据牌号不同，柴油可分为 10 号轻柴油样本 10 个和 0 号轻柴油样本 10 个。汽油可分为 90 号汽油样本 20 个、93 号汽油样本 18 个和 97 号汽油样本 10 个。因此，共计有 7 类石油产品样本。

2.2 设备

实验使用的拉曼光谱测量系统如图 2 所示，由 785 nm 半导体激光器、拉曼探头、CCD 背照式拉曼光谱仪以及连接光纤组成。石油产品样本放置在石英比色皿中。实验时，石英比色皿中的石油产品样本被激光器发出的 785 nm 波长的激光照射，能级产生变化，发出拉曼散射光。拉曼散射光通过光纤传输到光谱仪进行采样和 A/D 转换，最终转换结果被传输给计算机，从而获得石油产品样本的拉曼谱图。

2.3 谱图获取与预处理

采用上述的实验设备，对 7 类石油产品样本进行拉曼光谱采集，获得的谱图如图 3 所示。

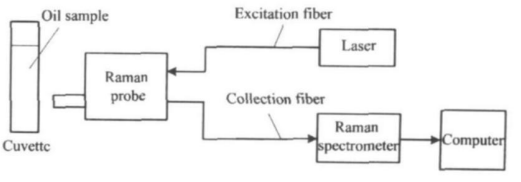

Fig 2 Raman measurement system

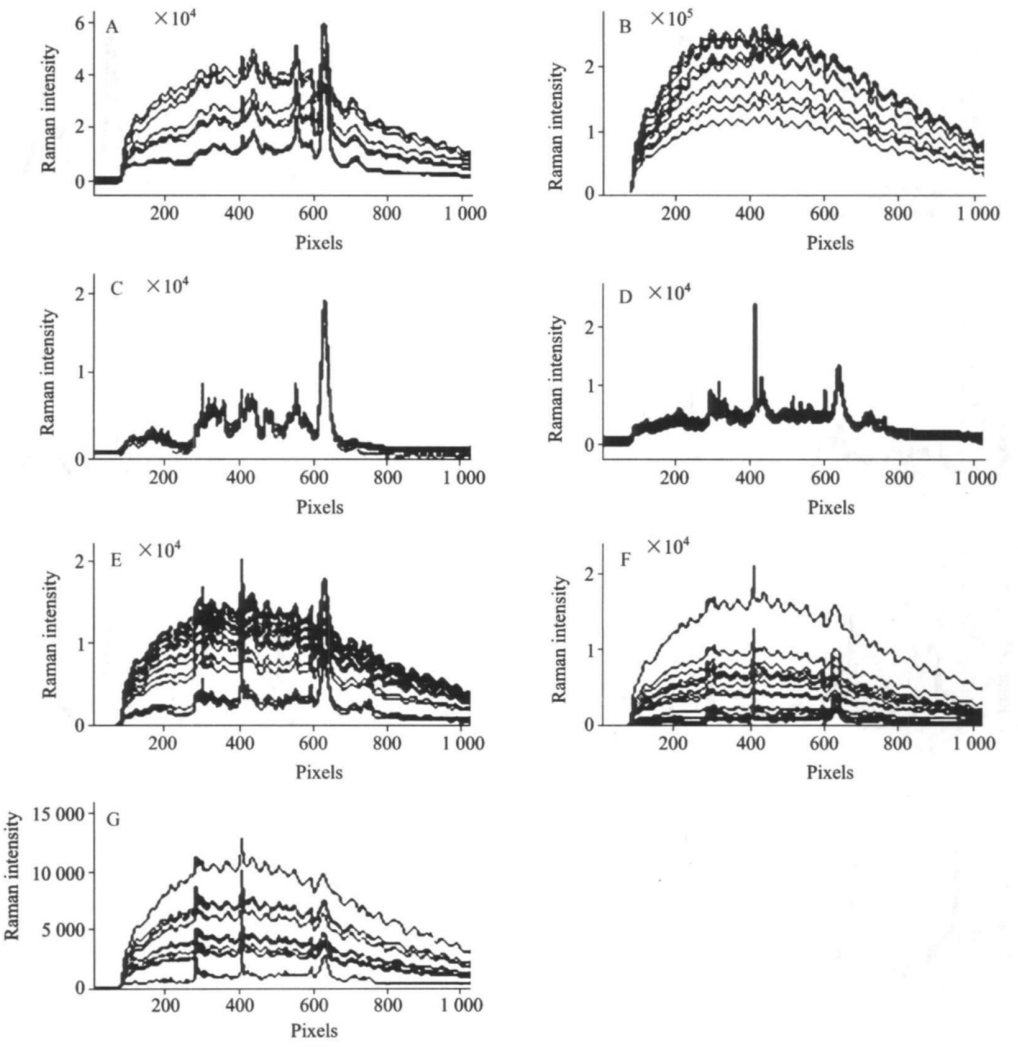

Fig 3　Original Raman spectra of petroleum samples

A: -10# Diesel; B: 0# Diesel; C: Naphtha; D: Methanol Gasoline; E: 90# Gasoline; F: 93# Gasoline; G: 97# Gasoline

对于谱图预处理，首先在去除干涉光时采用基于标准物质 SRM2241 的相对强度校正法[9]，由于谱图前段拉曼信号较弱，噪声相对较强，所以干涉光去除范围限制在 101~1 024像素点。其次，选择拉曼信号比较丰富的 271~ 780 像素点谱图段进行荧光背景的扣减，具体扣减算法采用 Lieber 等提出的迭代背景谱线拟合算法[10]。经过干涉光去除和荧光背景扣减步骤后，将拉曼谱图进行整数插值(将横坐标单位转换成拉曼位移，单位为：cm^{-1})，4 cm^{-1}为半窗宽的多项式平滑滤波和平均值归一化。所谓的平均值归一化指的是所有谱图均除以各自的拉曼光强的平均值，至此，预处理完成，谱图预处理的最终结果如图 4所示。

3　分类测试结果与讨论

为方便起见，将七类石油产品类别用字母代号表示，A: -10号轻柴油，B: 0 号轻柴油，C: 轻质石脑油，D: 甲醇汽油，E: 90 号汽油，F: 93 号汽油，G: 97 号汽油。依次将 96 个样本进行 1~ 96 的数字编号，其中 A 包含 1~ 10 号样本，B 包含 11~ 20 号样本，C 包含 21~ 30 号样本，D 包含 31~48 号样本，E 包含 49~ 68 号样本，F 包含 69~ 86 号样本，G 包含 87~ 96 号样本。

按照石油产品快速分类算法的步骤，对于 7 类石油产品在每类中随机抽出 3 个样本共计 21 个样本加入测试样本集，为检验方法对未知样本的判断能力，同时添加间二甲苯、邻二甲苯、对二甲苯和某个煤油样本的拉曼谱图进入测试样本集，其预处理完成的谱图如图 5 所示。7 类石油产品中剩余的所有样本为训练样本集。利用训练样本集对应的预处理完成的拉曼谱图和谱图所属类别构建模型知识库，计算每个类的类别特征拉曼谱图和类内阈值。最后，对测试样本集内样本进行分类判断。

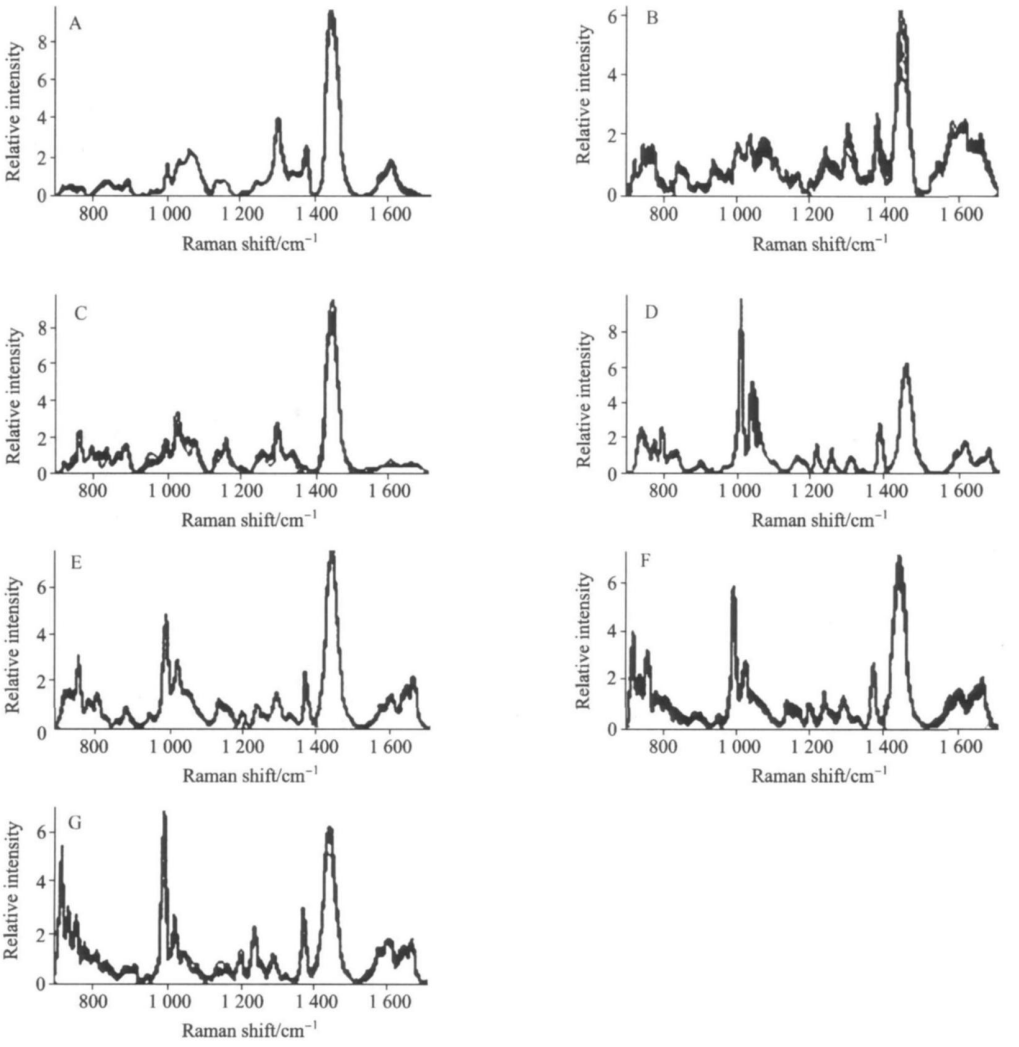

Fig 4 Preprocessed Raman spectra of petroleum samples

A: -10# Diesel; B: 0# Diesel; C: Naphtha; D: Methanol Gasoline; E: 90# Gasoline; F: 93# Gasoline; G: 97# Gasoline

共重复上述步骤 20 次，结果未发现一例错误分类情况，未知样本也全部被检出，其中第十次测试的计算中间结果如下。

Table 1 Distribution of test samples

Known sample class	Selected samples	Unknown sample class	Unknown sample property
-10# Diesel	1, 2, 7	M-xylene	Pure Substance
0# Diesel	15, 17, 19	O-xylene	Pure Substance
Naphtha	23, 25, 29	P-xylene	Pure Substance
Methanol Gasoline	31, 45, 48	Kerosene	Mixture
90# Gasoline	53, 61, 64		
93# Gasoline	72, 78, 80		
97# Gasoline	90, 93, 95		

挑选测试样本的结果如表 1 所示，模型知识库类内阈值计算结果如表 2 所示，表 3 中包含了测试样本谱图与模型知识库中不同类别特征谱图的相关系数，由表 3 可知：对于每个模型知识库已知类别的石油产品测试样本，其计算所得的最大相关系数对应的类内阈值均大于该最大相关系数。在判断类别结果中，没有出现错误分类的情况；对于每个未知类别的测试样本，其计算所得的最大相关系数均小于对应类别的类内阈值，因此均被判定为未知样本。

石油产品的分类测试结果表明，本文提出的基于拉曼光谱的石油产品快速分类算法，对模型知识库已知的石油产品样本具有良好的分类能力，在随机测试的条件下，未出现错误分类的情况；对于未知的样本，如某些纯物质，其他石油产品的谱图，该算法能有效将其辨别。可见，由线性相关系数体现谱图形状的类似，进而体现谱图对应物质化学组成的类似是可行的。当然，利用算法相对复杂的(soft independent modeling of class analogy, SIMCA)，PCA 或(support vector machine, SVM)等等分类技术，也可完成有效的分类，但所有复杂的分类技术都需人工给定分类界限的阈值范围，而该阈值的选择并没有实际的物理意义作为支撑，阈值选择也成为使用这些技术的关键和难点。本算法自动给定类内阈值，

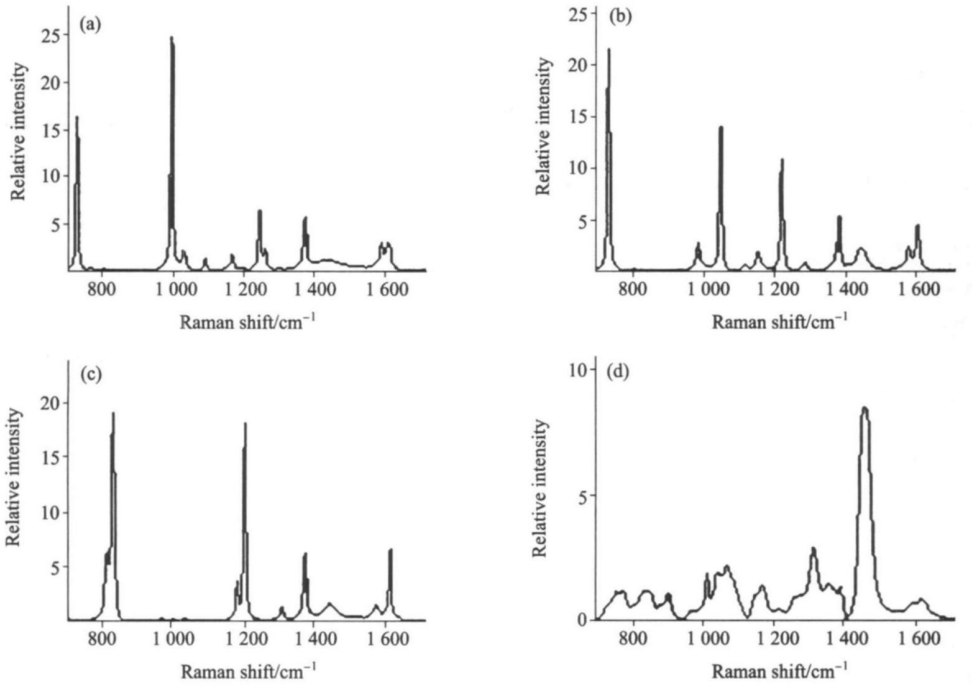

Fig 5 Preprocessed Raman spectra of unknown samples

(a): M-xylene; (b): O-xylene; (c): P-xylene; (d): Kerosene

Table 2 Intra-class threshold of each known class in the knowledge base

Known sample class in knowledge base	A	B	C	D	E	F	G
Intra-class threshold *R*min	0 991 5	0 935 9	0 957 3	0. 955 8	0 988 6	0 960 6	0 984 8

Table 3 Correlation coefficients between test samples and intra-class feature spectra

Test samples	Correlation coefficients							Classification results
	A	B	C	D	E	F	G	
1	**0 998 4**	0 802 1	0 942 3	0 619 6	0 859 8	0 813 0	0 713 2	A
2	**0 998 8**	0 836 5	0 927 5	0 621 0	0 857 3	0 814 5	0 720 9	A
7	**0 997 8**	0 797 2	0 943 4	0 617 4	0 859 5	0 812 4	0 711 7	A
15	0 934 8	**0 967 3**	0 853 2	0 660 6	0 824 3	0 802 3	0 734 3	B
17	0 903 9	**0 982 8**	0 818 1	0 654 3	0 802 5	0 786 3	0 726 9	B
19	0 869 4	**0 994 9**	0 785 6	0 651 4	0 790 8	0 779 7	0 726 2	B
23	0 936 0	0 738 9	**0 999 0**	0 693 0	0 908 1	0 860 7	0 752 5	C
25	0 929 4	0 734 8	**0 997 5**	0 680 6	0 901 0	0 854 8	0 746 1	C
29	0 930 1	0 737 5	**0 998 4**	0 708 6	0 915 8	0 869 6	0 762 5	C
31	0 582 9	0 604 7	0 643 3	**0 986 9**	0 809 0	0 860 1	0 903 5	D
45	0 634 3	0 651 0	0 696 6	**0 999 3**	0 843 3	0 883 8	0 909 8	D
48	0 619 6	0 629 8	0 695 8	**0 988 3**	0 823 3	0 855 8	0 871 8	D
53	0 866 7	0 777 3	0 914 0	0 821 6	**0 996 1**	0 975 6	0 898 6	E
61	0 858 0	0 765 4	0 906 0	0 942 3	**0 999 7**	0 983 1	0 913 1	E
64	0 858 6	0 771 1	0 905 6	0 839 3	**0 999 6**	0 982 5	0 911 5	E
72	0 804 1	0 781 1	0 845 6	0 886 0	0 969 0	**0 987 4**	0 960 1	F
78	0 822 5	0 785 7	0 865 6	0 858 8	0 969 2	**0 987 4**	0 954 7	F
80	0 806 6	0 764 1	0 855 3	0 892 7	0 974 4	**0 996 8**	0 972 8	F
90	0 718 1	0 732 9	0 757 2	0 882 2	0 893 1	0 955 1	**0 987 4**	G
93	0 733 9	0 723 1	0 767 0	0 902 9	0 922 8	0 974 5	**0 997 4**	G
95	0 734 8	0 711 4	0 767 4	0 874 5	0 909 0	0 968 5	**0 994 9**	G
M-xylene	0 219 7	0 275 9	0 357 5	0 181 8	0 078 4	0 059 2	**0 460 3**	unknown

续表 3

O-xylene	0.116 2	0.132 1	0.197 5	**0.211 6**	0.113 7	0.068 1	0.184 8	unknown
P-xylene	0.016 5	0.015 2	0.014 7	-0.014 6	**0.032 4**	-0.022 8	0.031 3	unknown
Kerosene	0.742 5	0.687 8	0.583 0	0.649 7	**0.875 5**	0.812 8	0.518 6	unknown

无需人工判断，不仅物理意义清晰明确，且算法实现及其简单，便于实际应用。

类内阈值体现了同类石油产品样本拉曼谱图的最大差异性，若计算出的最大相关系数小于类内阈值，则证明目前差异已经大于同类石油产品样本的最大差异。因此，该样本被划分为不同类，即判断为未知样本。从分类测试结果看，90号样本的最大相关系数为 0.987 4，仅比对应的类内阈值 0.984 8 略大一些，存在将其判断为未知样本的风险。所以，类内阈值的设置还有进一步优化的空间。

4 结 论

本文提出了一种基于拉曼光谱的石油产品快速分类方法，在利用石油产品训练样本构建模型知识库的基础上，依照测试样本谱图和类别特征谱图的线性相关度决定待测样本的类别。实验结果表明：该方法能够有效辨别模型知识库已知的石油产品样本，并成功判断出未知样本。该方法算法结构和实现简单，物理意义清晰明确，可有效应用于工业油品传输和快速分类等领域。

References

[1] LU Wan-zhen(陆婉珍). Modern Near Infrared Spectroscopy Analytical Technology, Second Edition(现代近红外光谱分析技术，第 2 版). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2007.

[2] YANG Xu-gang, WU Qi-lin(杨序纲, 吴琪琳). Raman Spectroscopy Analysis and Application(拉曼光谱的分析与应用). Beijing: National Defense Industry Press(北京: 国防工业出版社), 2008.

[3] Kim M, Lee Y H, Han C G. Computers and Chemical Engineering, 2000, 24(2-7): 513.

[4] Balabin R M, Safieva R Z. Fuel, 2008, 87(7): 1096.

[5] Balabin R M, Safieva R Z, Lomakina E I. Analytica Chimica Acta, 2010, 671(1-2): 27.

[6] ZHANG Qi-ke, DAI Lian-kui(张其可, 戴连奎). Control and Instruments in Chemical Industry(化工自动化及仪表), 2005, 32(4): 53.

[7] Flecher P E, Welch W T, Albin, S, et al. Spectrochimica Acta Part A, 1997, 53(2): 199.

[8] Cooper J B, Wise K L, Welch W T, et al. Applied Spectroscopy, 1997, 51(11): 1613.

[9] Choquette S J, Etz E S, Hurst W S, et al. Applied Spectroscopy, 2007, 61(2): 117.

[10] Lieber C A, Mahadevan-Jansen A. Applied Spectroscopy, 2003, 57(11): 1363.

A Fast Classification Method for Petroleum Products Based on the Raman Spectroscopy

LI Sheng, DAI Lian-kui*

State Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China

Abstract A fast and effective method for classification of petroleum products based on Raman spectroscopy is proposed. A knowledge base composed by Raman spectra of training samples, intra-class feature spectra and intra-class thresholds of all classes was firstly established. Then, correlation coefficients between the test sample and the intra-class feature spectra were calculated. If the maximal correlation coefficient of the test sample is larger than or equal to the corresponding intra-class threshold, the test sample is determined to belong to the corresponding class. For 96 petroleum product samples belonging to 7 classes and 4 unknown samples, the experimental results show that this method can accurately classify known test samples and can also find the unknown test samples. This method costs little calculation time and human interference. Moreover, it can be easily implemented in the practical application.

Keywords Raman spectroscopy; Petroleum products; Fast classification; Correlation analysis

* Corresponding author

(Received Jan. 16, 2011; accepted May 20, 2011)