对拉曼光谱信号分类的模式识别 [1]

郭 平 卢汉清 * 杜为民 **

(北京师范大学信息科学学院 北京 100875)
*(中国科学院自动化所模式识别国家重点实验室 北京 100080)
**(北京大学物理系 100871)

摘 要 该文提出了应用重整化的高斯分类器对拉曼光谱信号识别. 把每条光谱信号看作是高维矢量空间的一个点, 建立统计模型后采用贝叶斯定则进行概率分类. 计算机实验结果表明识别正确率可达到 99.81% .

关键词 拉曼光谱, 模式识别, 分类, 贝叶斯定则, 重整化, 甄别分析
中图号 TP391.4

1 引 言

拉曼光谱是一种测量分子振动光谱的方法, 它和红外光谱的物理机制不同, 但它们是互为补充的. 根据对电磁辐射的响应, 分子振动光谱分为红外活性和拉曼活性两类. 拉曼光谱适用于测量红外活性弱的分子, 有清楚的特征峰, 没有通常红外光谱中由于高阶振动模式的吸收所产生的干扰, 有利于鉴定样品的成分. 它有利于测量水溶液中的样品而没有在红外光谱中的来自于水的干扰, 所以拉曼光谱是一种很好的测量技术. 用在线拉曼光谱测量的方法进行反应监测, 不需取样分离就可远程监测化学反应中各组分的拉曼光谱随时间变化的关系曲线. 对乙醇、乙酸合成乙酸乙酯的化学反应过程中, 用实时在线拉曼光谱的方法进行跟踪测量, 得到了在反应过程中反应物和生成物的拉曼光谱随时间的变化的数据. 本文利用所得到的拉曼光谱信号, 基于统计模式识别技术, 采用重整化的高斯分类器, 对乙醇、乙酸, 以及乙醇、乙酸和乙酸乙酯混合物这 3 类样品进行识别研究. 目的是建立全自动计算机数据分析与处理的模型, 为其工业生产过程自动化打下基础.

2 分类技术

2.1 贝叶斯概率分类与甄别分析的关系

在假定样本点是服从独立全同分布 (independent, identical distribution, i.i.d) 时, 根据样本, 我们可估计待识别样本的后验概率:

$$p(j|x)=\frac{\alpha_j G(x,m_j,\Sigma_j)}{\sum_{l=1}^{k}\alpha_l G(x,m_l,\Sigma_l)} \tag{1}$$

其中 $G(x,m_j,\Sigma_j)=\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left[-\frac{1}{2}(x-m_j)^T\Sigma_j^{-1}(x-m_j)\right]$ 是多变量高斯概率密度分布的一般表示。式中 x 代表随机矢量, d 是随机矢量 x 的维数。 α_j 是混合权重参数 (或先验概率), 要求 $\alpha_j\geq 0, \sum_{j=1}^{k}\alpha_j=1$, m_j 是均值矢量, 而 Σ_j 是协方差矩阵。

后验概率函数 (1) 式表示样本 x 属于类别 j 的概率。对给定的 x_i 样本点, 利用贝叶斯定则, $j^*=\arg\max_j p(j|x_i)$, 把样本 x_i 划分到类别 j^* 中, 这个过程称为贝叶斯概率分类。

[1] 2000-08-23 收到, 2001-04-23 定稿

实际上，如果我们对 (1) 式取自然对数，得到：

$$\ln p(j|x_i) = \ln \alpha_j G(x_i, m_j, \Sigma_j) - \ln \sum_{l=1}^{k} \alpha_l G(x_i, m_l, \Sigma_l) \tag{2}$$

由于自然对数 $\ln(\bullet)$ 是凸函数，最大化 $p(j|x_i)$ 等价于最大化 $\ln(p(j|x_i)$ 。而且上式右边第二项对每一类都相同，忽略后对分类没影响。(2) 式右边第一项可以写为

$$\ln \alpha_j G(x_i, m_j, \Sigma_j) = \ln \alpha_j - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x - m_j)^T \Sigma_j^{-1}(x - m_j) \tag{3}$$

在忽略类别的公共因子和常数后，分类准则可写为

$$j^* = \arg\min_j d_j(x_i) \tag{4}$$

和

$$d_j(x_i) = (x - m_j)^T \Sigma_j^{-1}(x - m_j) + \ln|\Sigma_j| - 2\ln\alpha_j \tag{5}$$

在文献 [1] 中 (5) 式被称为甄别得分 (discriminant score) 。更进一步，在每一类的先验概率都相同时，可抛弃 $2\ln\alpha_j$ 项，(5) 式则成为甄别函数。从上面建立的关系来看，在使用高斯分类器时，贝叶斯概率分类与甄别分析实际上是等价的。

2.2 重整化的甄别分析

要进行贝叶斯概率分类，首先要根据已知的样本建立高斯分类器，这需要估计均值与协方差。通常采用的样本均值估计是 [2]

$$m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i \tag{6}$$

最大似然协方差估计为

$$\Sigma_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(x - m_j)(x - m_j)^T \tag{7}$$

或无偏样本协方差估计为

$$\Sigma_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x - m_j)(x - m_j)^T \tag{8}$$

其中 n_j 是每类样本的数目。采用上面的样本协方差估计与分类准则 (4) 式和 (5) 式，即为通常所采用的二次甄别分析 (Quadratic Discriminant Analysis, QDA)[1] 。

当待分析的样本数目 n_j 小于样本矢量维数 d 时，采用上述的样本协方差估计得到的矩阵变成奇异矩阵。这时问题成为不适定 (ill-posed) 问题，用 QDA 分类得到的结果是不可靠的。在高维小样本情况时，有两种方法可以采用，降维与重整化 [1] 。线性甄别分析 (Linear Discriminant Analysis, LDA) 则是重整化方法之一，

$$\Sigma = \frac{1}{N}\sum_{j=1}^{k} n_j \Sigma_j \tag{9}$$

N 是样本的总数。用公共矩阵 Σ 代替每类的单个矩阵，在某些情况下可显著改善分类器的识别正确率。重整化甄别分析 (Regularized Discriminant Analysis, RDA)[3] 则是另外一种方法

$$\Sigma_j(\lambda,\gamma)=(1-\gamma)\Sigma_j(\lambda)+\gamma\frac{\mathrm{Trace}(\Sigma_j(\lambda))}{d}\boldsymbol{I}_d \tag{10}$$

其中

$$\Sigma_j(\lambda)=\frac{(1-\lambda)n_j\Sigma_j+\lambda N\Sigma}{(1-\lambda)n_j+\lambda N} \tag{11}$$

式中 $\boldsymbol{I}_d$ 是 $d\times d$ 维的单位矩阵，λ 和 γ(取值范围限制在 0 与 1 之间) 是按照最大 leave-one-out 分类精确度来选择的重整化参数。λ 控制 Σ_j 朝 Σ 收缩的量度，由于量 $\mathrm{Trace}(\Sigma_j(\lambda))/d$ 等价于 $\Sigma_j(\lambda)$ 本征值的平均，则可认为 γ 控制本征值向相等性收缩的程度。

由于在 RDA 中重整化参数需采用交叉检验统计技术来估计，计算工作量较大，我们还采用了库勒巴克-雷伯勒信息量度 (Kullback-Leibler Information Measure, KLIM) 重整化方法 [4]。

$$\Sigma_j(h)=h\boldsymbol{I}_d+\Sigma_j \tag{12}$$

其中 h 是非参数核密度函数中的平滑参数 [2]，在这里起重整化参数的作用，可用下式估计 [4]

$$h=\frac{d}{2J_r(x_i,\Theta)},\quad J_r(x_i,\Theta)=-\frac{1}{2N}\sum_{i=1}^{N}\mathrm{Trace}\left\{\nabla_x^2\ln\left[\sum_{l=1}^{k}\alpha_l G(x_i,m_l,\Sigma_l)\right]\right\} \tag{13}$$

在计算重整化参数的问题上，KLIM 重整化方法的计算工作量要比 RDA 重整化方法中交叉检验统计技术小好多。

3 实验结果与讨论

3.1 数据获取

实验采用的 CCD 探头横向具有 1340 道，测量的光谱范围为 320–1640 cm^{-1}。纯乙醇、乙酸的静态拉曼光谱，除了本底高斯噪音外基本不变。对这两种物质的拉曼光谱分别测量了 3 次与 5 次。而对乙醇、乙酸合成乙酸乙酯的化学反应过程，实验中每隔一定的时间进行一次拉曼光谱的测量，就可以得到拉曼光谱的变化规律。在反应开始时的光谱应是乙醇与乙酸两者的叠加，随着反应的进行，这两种成分应逐渐减少，而乙酸乙酯逐渐增加。因为乙醇与乙酸不可能全部反应掉，最终得到的是 3 种物质混合的拉曼光谱曲线。我们从合成过程接近尾声时采集的众多数据中随机选取了 29 条光谱作为识别的样本。

图 1(a)，(b) 和 (c) 分别是纯乙醇、乙酸以及乙醇、乙酸和乙酸乙酯混合物的拉曼光谱。乙醇的拉曼峰与乙酸的拉曼峰相距很近，在 $882\mathrm{cm}^{-1}$ 处重叠。乙酸的拉曼峰与乙酸乙酯的拉曼峰在 620 cm^{-1} 附近略有重叠，但仍有可以分辨反应物与生成物的特征峰。$847\mathrm{cm}^{-1}$ 处是乙酸乙酯的特征峰。在反应过程进行到一定程度时，乙酸乙酯的拉曼峰增加到极大值，乙醇、乙酸的拉曼峰减小到极小值，反应基本稳定，乙酸乙酯和乙醇、乙酸的光谱强度基本不变，有少许波动。

3.2 数据预处理

从上可见，每个原始数据是 1340 道，如果直接使用，则维数太高，样本数目太少，分类效果不可能好。我们采用均匀抽样的方法，每隔 10 道抽取一个数据作为新的矢量的一维。用这种数据预处理技术，构造的新矢量是 134 维。一个原始样本分解成为 10 个新的样本。这样数据样本数目成为 370 个。

一般来讲，光谱特征谱线的强度与反应物和生成物的浓度成线性关系，原理上可得到反应物和生成物的浓度变化规律，进而对化工生产进行实时控制。但由于光谱强度是任意单位，只

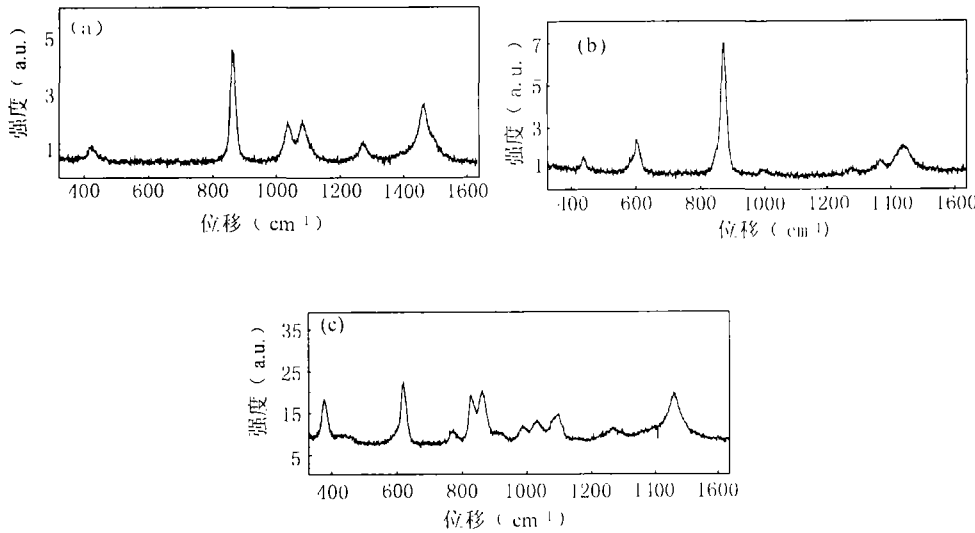

图 1 (a), (b) 和 (c) 分别是纯乙醇、乙酸以及乙醇、乙酸和乙酸乙酯混合物的拉曼光谱

有相对强度才是具有意义的。不失一般性，我们对光谱强度进行了归一化处理，即把强度变换到 0~1 范围之内。

3.3 识别结果与讨论

在对拉曼光谱的识别过程中，我们把乙醇、乙酸以及乙醇、乙酸和乙酸乙酯混合物看作是三类物质。从预处理后的每一类物质的拉曼光谱数据中随机抽取 20 个样本作为训练样本，用训练样本来估计分类器的均值与协方差矩阵。剩余的 310 个样本则用于检验分类的准确率。由于每类的训练样本是 20，而维数是 134，该问题是不适定问题。

在识别实验中，对两个重整化参数 λ 和 γ 采用较为粗糙的网格取样，$(0, 0.25, 0.5, 0.75, 1.0)$，这样得到 25 个重整化参数的数据点。实验重复了 26 次，识别结果的平均精确率如表 1 所示。在表 1 中平均精确率用百分比表示，括号里的值表示标准偏差，而 NS 表示协方差矩阵是奇异的，这种情况下得不到可靠的结果。

表 1 对拉曼光谱信号分类的平均精确率

分类技术	QDA	LDA	RDA	KLIM
识别率 %	NS	NS	99.27(0.43)	99.81(0.28)

本文采用重整化的高斯分类器，不同于通常的光谱识别方法。如果用特征蜂的位移，强度来识别，则由于强度的任意性，特征蜂的重叠性造成模式识别的困难。而我们在本文中采用的是整条光谱曲线，每道数据表示输入矢量在多维空间的某一维上的投影值，原来在低维空间不可分的样本在高维空间则可分辨。我们利用了整个光谱的特征，按所有维计算概率，采用贝叶斯定则来分类。可以看出，输入变量包含了每一道数据，这是一种集体决策过程，减低了误分类的风险。

4 结论与进一步的工作

本文基于统计模式识别技术，提出了应用重整化的高斯分类器，对拉曼光谱信号实施计算机智能处理。我们把每条光谱信号看作是高维矢量空间的一个点，所有的样本构成样本空间。在假定样本点是服从 i.i.d 后，建立了统计模型并采用贝叶斯定则进行概率分类。样本点服从 i.i.d

是统计方法处理问题的一个基本假定，大多数情况是满足这个条件的．计算机模拟实验结果表明识别正确率可达到 99.81%，这说明我们提出的方法和采用的策略是非常成功的．

需要指出的是我们在本文中采用了重整化方法，并得到了相当满意的结果．但是在实际应用中所遇到的问题是多样的，例如有时所获取信号的信噪比不太好，先验概率的假定与实际数据分布相差太远等原因，有时用统计方法可能得到的识别正确率不是太理想．我们知道基于统计技术的模式识别分类并不是唯一的，实际上可以有好多方法来解决分类问题，例如前馈神经网络分类器等[5]．即使对不适定问题，还可以采用降维方法把问题化解为适定问题．我们进一步的工作准备结合主分量分析与高斯分类器，探讨其对拉曼光谱信号识别的正确率，并与重整化的高斯分类器进行比较．

参　考　文　献

[1] Stefan Aeberhard, Danny Coomans and Olivier de Vel, Comparative analysis of statistical pattern recognition methods in high dimensional settings, Pattern Recognition, 1994, 27(8), 1065–1077.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, second edition, Boston, Academic Press, 1990, chapter 2.

[3] J. H. Friedman, Regularized discriminant analysis, J. of American Statistics Association, 1989, 84(405), 165–175.

[4] Ping Guo, Michael R. Lyu, Classification for high-dimension small-sample data sets based on kullback-leibler information measure, Proc. of the Int. Conference on Artificial Intelligence (IC-AI'2000), Las Vegas, Nevada, USA: CSREA Press, 2000, 1187–1193

[5] C. M. Bishop, Neural Networks for Pattern Recognition, 1995, Oxford, Oxford University Press, 1995, 80–90.

PATTERN RECOGNITION FOR THE CLASSIFICATION OF RAMAN SPECTROSCOPY SIGNALS

Guo Ping　　Lu Hanqing*　　Du Weimin**

(*College of Information Science, Beijing Normal University, Beijing* 100875, *China*)

*(*National Laboratory of Pattern Recognition, CAS, Beijing* 100080, *China*)

**(*Dept. of Physics, Peking University, Beijing* 100871, *China*)

Abstract　This paper presents a study on application of regularized Gaussian classifier to Raman spectroscopy signals pattern recognition. Each Raman spectrum is treated as a point in high dimensional vector space. For given samples, the statistical model is used and Bayes decision is adopted to classify then according to their maximum posterior probabilities. Computer experiments show that the classification accuracy is obtained as high as 99.81%.

Key words　Raman spectroscopy, Pattern recognition, Classification, Bayes decision, Regularization, Discriminant analysis

郭　平：　男，1957 年生，教授，研究方向：光谱分析，神经网络，模式识别，软件可靠性，智能信息处理．

卢汉清：　男，1961 年生，研究员，博士生导师，图像处理和图形学研究组负责人．目前感兴趣的研究工作包括：图像理解及应用、多媒体技术及信息系统、医学图像处理等．

杜为民：　男，1952 年生，1995 年于美国纽约市立大学获博士学位，现任北京大学物理系副教授，从事激光光谱及凝聚态物质的光谱学性质的研究．