基于集成特征的拉曼光谱谱库匹配方法

刘铭晖[1,2]，董作人[1*]，辛国锋[1]，孙延光[1]，瞿荣辉[1]，魏芳[1]，殷磊[3]

[1]中国科学院上海光学精密机械研究所空间激光信息传输与探测技术重点实验室，上海 201800；

[2]中国科学院大学，北京 100049；

[3]南京简智仪器设备有限公司，江苏 南京 210038

摘要 基于谱库的匹配识别是应用拉曼光谱进行物质成分鉴别的关键，会直接影响匹配结果的准确性。在谱库匹配中，尤其是针对混合物的光谱，利用单一的匹配特征无法全面反映被测样本光谱与谱库光谱的相似性，光谱匹配识别时需要综合考虑多种匹配特征。采用逻辑回归数学模型融合谱峰匹配系数、非负最小二乘匹配系数以及夹角余弦匹配系数，提出了一种新的光谱集成匹配方法。该方法既考虑了被测样品的谱峰信息，又考虑了其全谱信息。基于 20 种氨基酸混合物拉曼光谱谱库匹配的实验结果表明：所提光谱集成匹配方法具有更低的误判率。

关键词 光谱学；拉曼光谱；谱库匹配；逻辑回归模型

中图分类号 O433 **文献标识码** A **doi**：10.3788/CJL201946.0111002

Raman Spectrum Library Matching Method Based on Integrated Features

Liu Minghui[1,2], Dong Zuoren[1*], Xin Guofeng[1], Sun Yanguang[1], Qu Ronghui[1], Wei Fang[1], Yin Lei[3]

[1]*Key Laboratory of Space Laser Communication and Detection Technology, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai* 201800, *China*;

[2]*University of Chinese Academy of Sciences, Beijing* 100049, *China*;

[3]*Nanjing Simple & Smart Instrument and Equipment Co., Ltd., Nanjing, Jiangsu* 210038, *China*

Abstract Library-based matching recognition is the key to the application of Raman spectroscopy for material composition identification, which directly affects the accuracy of matching results. For the library matching, especially for the mixture spectrum, the single matching feature can not fully reflect the similarity between the measured sample spectrum and the spectral spectrum in the library. The spectral matching needs to comprehensively consider multiple matching features. In this paper, a new spectral integration matching method is proposed by using the logistic regression mathematical model to fuse the peak matching coefficient, the non-negative least squares matching coefficient and the cosine matching coefficient. The new method takes into account both the spectral peak information and full spectrum information of the sample. The experiment results based on the Raman spectroscopy library matching of 20 kinds of amino acid mixtures show that the spectral integration matching method has lower false positive rate.

Key words spectroscopy; Raman spectroscopy; spectrum library matching; logical regression model

OCIS codes 300.6450; 170.5660; 200.4560

1 引 言

拉曼光谱是一种分子散射光谱，由激光照射在物质上发生的非弹性散射产生，谱峰的位置和强度直接反映了物质的成分及含量信息，因此拉曼光谱也被称为物质的指纹图谱[1]。拉曼光谱具有快速、无损、样品无需预处理以及可在线分析等优点，已被广泛应用于食品、材料、医疗等领域[2-3]，用于物质成分的判定以及快速分类。目前，基于拉曼光谱分析的模式识别方法主要分为三类：无监督识别方法、有

收稿日期：2018-09-07；**修回日期**：2018-10-09；**录用日期**：2018-10-17

基金项目：国家自然科学基金(61535014,61775225)、上海市科技成果转化和产业化项目(18511104500)

* **E-mail**：zrdong@siom.ac.cn

监督识别方法、光谱库检索方法[4]。前两种方法根据样本类别进行定性分析；光谱库检索方法根据待测样本的光谱从已构建好的光谱库中检索出与被测样本相似度最高的一个或多个样本，从而实现定性甚至定量分析。其中，相似度常使用相关系数、夹角余弦、欧氏距离和光谱信息散度等方法计算得到。然而，目前的光谱检索方法多用于纯净化合物的光谱识别，混合物中物质的定性分析依然是研究人员面临的难题和挑战[5-6]。

随着检测样本的组成趋于复杂，光谱检索算法越来越受到研究人员的关注[7]，一些新的检索算法和检索策略使光谱检索的准确性和可靠性得到了显著提高。Zhang 等[8]结合小波变换寻峰和逆检索算法对甲醇、乙醇、乙腈液体混合物以及聚丙烯酰胺、乙酸钠、碳酸钠粉末混合物进行识别，并用非负最小二乘(NNLS)算法进行筛选，识别的准确率高于相关系数的识别率。Gawinkowski 等[9]将 Canberra 距离作为相似度指标，结合加法模型对固态氨基酸混合物进行识别，识别的正确率明显优于偏最小二乘法(PLS)以及非负最小二乘判别法，但该算法的运算速度较慢，不适宜进行快速检测。彭颖等[10]将拉曼谱峰信息转换到小波域空间，利用小波域空间谱峰信息和逆检索算法定义了新的反向搜索匹配系数，并对食品中的色素进行谱库匹配和定性分析；与传统的质量系数(HQI)匹配算法相比，该算法的匹配准确率有较大提升，但其匹配准确率仍受限于寻峰的准确度。孔祥兵等[11]将欧氏距离、相关系数和光谱信息散度三种方法进行集成后对高光谱遥感影像进行识别，该方法具有更强的光谱判别力和更小的光谱识别不确定性。目前，基于谱库识别方法的研究主要集中在两个方面：一是基于传统的相关系数和夹角余弦法等方法，根据被分析光谱的特点构建新的光谱相似度；二是通过改进检索策略来获得更加准确的结果。

只采用一种谱库光谱匹配方法得到的结果可能不够稳定，而采用集成策略被认为是解决这一问题的有效手段[12]。集成策略的基本思想是采用多种匹配算法分别建立识别规则，得到各自的判别系数，然后同时对待测样本的光谱进行鉴别分析，通过加权或其他数学模型对判别系数进行融合，得到最终的判别结果。该检索策略降低了检索结果对单一匹配算法的依赖性，提高了检索结果的稳定性。本研究通过线性逻辑回归融合谱峰匹配系数(PMC)、非负最小二乘匹配系数、夹角余弦匹配系数计算得到最终的匹配系数。所提匹配算法综合考虑了光谱谱峰信息和全谱信息，相较于单一的匹配算法具有更低的误判率。

2　基本原理

在谱库中对拉曼光谱进行匹配时，综合考虑谱峰匹配、非负最小二乘匹配、夹角余弦匹配这三种匹配系数，即综合考虑谱峰信息与全光谱信息对光谱进行匹配识别，这便是本课题组所提集成匹配算法的主要思想。下面将讨论谱峰匹配系数、非负最小二乘匹配系数、夹角余弦匹配系数的表达式，最后给出集成匹配系数的数学模型。

2.1　光谱匹配系数的定义

2.1.1　谱峰匹配系数

针对谱库中物质的拉曼光谱，使用多尺度谱峰检测(MSPD)算法进行谱峰检测[13]，采用基于 Voigt 函数拟合的谱峰判别算法进行谱峰判定[14]，得到每一个谱峰的拟合峰高 a_i，定义每一个谱峰的权重 $\omega_{\mathrm{peak}}(i)$ 为

$$\omega_{\mathrm{peak}}(i)=\frac{a_i}{\sum_{k=1}^{K}a_i},\tag{1}$$

式中：K 为谱峰总数。从定义中可得出，谱峰的峰高越大，其权重越大。谱峰匹配系数的定义为，谱库光谱中物质光谱与被测样本光谱重合的谱峰的权重之和：

$$C_{\mathrm{PMC}}=\sum_{m=1}^{M}\omega_{\mathrm{peak}}(m),\tag{2}$$

式中：M 为重合谱峰的个数。谱峰匹配系数取决于被测样本光谱与谱库光谱谱峰的重叠程度，重叠的谱峰越多且谱峰高度越大，谱峰匹配系数就越大。谱峰匹配系数的取值范围为 0～1。

2.1.2　非负最小二乘匹配系数

在完成谱库谱峰检索后就可以得到谱库中与被测样本有重叠谱峰的物质。依据混合物光谱的叠加模型就可以认定被测样本即由这些物质中的一种或者几种组合而成，这些物质即组成了一个二级谱库。由于组成被测样本的物质的含量必须大于 0，因此，定义二级谱库中每种物质的非负最小二乘匹配系数 b_i 满足

$$\min\|\boldsymbol{y}-\boldsymbol{X}\cdot\boldsymbol{b}\|,b_i\geqslant 0,\tag{3}$$

式中：$\boldsymbol{y}$ 为被测样本强度归一化拉曼光谱；$\boldsymbol{X}$ 为由二级谱库中物质的强度归一化拉曼光谱组成的矩阵。该非负最小二乘问题可以由非负最小二乘算法进行

计算[15]，该算法由 Lawson 和 Hanson 提出，他们还证明了该算法具有有限的收敛。定义非负最小二乘匹配系数表达式为

$$C_{\mathrm{NMC}}=b_i。\tag{4}$$

2.1.3　夹角余弦匹配系数

夹角余弦用于评价两个光谱的相似程度。光谱 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 之间的夹角余弦匹配系数表达式为

$$C_{\mathrm{AMC}}=\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}|\cdot|\boldsymbol{y}|},\tag{5}$$

夹角越小，说明两个样本在模式空间中就靠得越近，相似性就越大。若两个光谱完全相同，则 $\cos(\boldsymbol{x},\boldsymbol{y})=1$；若两光谱完全不同，则 $\cos(\boldsymbol{x},\boldsymbol{y})=0$。夹角余弦是谱库检索最常用的方法[12]，其重点强调的是谱图之间整体的相似性。

2.1.4　集成匹配系数

谱峰匹配系数是通过特征峰信息对被测样本光谱与谱库光谱进行匹配的，反映了被测样本光谱的局部特征信息与谱库光谱的匹配程度；非负最小二乘匹配系数和夹角余弦匹配系数反映了被测样本光谱的全谱信息与谱库光谱的匹配程度。集成匹配系数指根据所计算出的三种匹配系数，通过数学模型融合得到的二级谱库中的物质可能是被测样品组分的概率。本研究所采用的数学模型为逻辑回归模型。

逻辑回归模型是线性模型的一种特殊形式，是一种概率统计分类模型，用于估计某个事件发生的概率。在线性模型中，假定因变量 z 可以被近似地表示为以 $\boldsymbol{x}$ 为因变量的线性函数：

$$z=\boldsymbol{\theta}^{\mathrm{T}}\cdot\boldsymbol{x}=\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_3+\cdots,\tag{6}$$

式中：$\boldsymbol{\theta}$ 为经过训练得到的权重系数向量；$\theta_i(i=0,1,2,\cdots)$ 即为因变量的权重系数。在逻辑回归模型中，线性函数为 sigmoid 函数：

$$h_\theta(\boldsymbol{x})=g(\boldsymbol{\theta}^{\mathrm{T}}\cdot\boldsymbol{x})=\frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\cdot\boldsymbol{x})}。\tag{7}$$

对于二分类问题，假设概率

$$P(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})=[h_\theta(\boldsymbol{x})]^z[1-h_\theta(\boldsymbol{x})]^{1-z},z\in\{0,1\}。\tag{8}$$

逻辑回归问题参数的更新规则采用梯度下降法，通过不断计算损失函数关于权重系数的梯度，并利用梯度的负方向为损失函数下降速度最快的方向为准则更新权重系数，使损失函数随着梯度的更新而不断下降。权重系数的更新方式为

$$\boldsymbol{\theta}^t=\boldsymbol{\theta}^{t-1}-\alpha\cdot\boldsymbol{x}[g(\boldsymbol{\theta}^{t-1}\cdot\boldsymbol{x})-\boldsymbol{y}],\tag{9}$$

式中：t 为迭代次数；α 为学习率。利用梯度下降法得到最优的权重系数后，就可以利用 $h_\theta(\boldsymbol{x})$ 来推测二级谱库中物质存在于被测样本中的概率。定义集成匹配光谱系数的表达式为

$$C_{\mathrm{IMC}}=h_\theta(\boldsymbol{x})=\frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\cdot\boldsymbol{X})},\tag{10}$$

式中：$\boldsymbol{X}$ 为匹配系数向量$[1,x_1,x_2,x_3]$，其中的 x_1 为谱峰匹配系数，x_2 为非负最小二乘匹配系数，x_3 为夹角余弦匹配系数。

2.2　模型评价指标

为了验证谱库匹配模型的判别效果，采用误判率和准确率来衡量匹配模型的准确性。误判率和准确率的表达式为

$$\begin{cases}T_{\mathrm{PR}}=\dfrac{V_{\mathrm{TP}}}{V_{\mathrm{TP}}+V_{\mathrm{FN}}}=\dfrac{V_{\mathrm{TP}}}{P},\\[2ex] F_{\mathrm{DR}}=\dfrac{V_{\mathrm{FP}}}{V_{\mathrm{FP}}+V_{\mathrm{TP}}},\end{cases}\tag{11}$$

式中：V_{TP} 为正确判断被测样本成分的个数；V_{FN} 为存在于被测样本中但没有匹配到的成分的个数；V_{FP} 为不存在于被测样本中而被误判为匹配到的成分的个数。模型评价指标的参数说明参见表 1。当对某一匹配模型进行评估时，在 T_{PR} 相同的情况下，F_{DR} 值越小说明模型的匹配判别性能越好。

表 1　模型评价指标参数说明

Table 1　Model evaluation index parameter description

Example	Judged as a positive example	Judged as a negative example
Positive example	V_{TP}	V_{FN}
Negative example	V_{FP}	

3　实验介绍与结果分析

3.1　样品与仪器

本研究讨论的混合物由 20 种氨基酸分析纯样（合肥博美生物科技有限公司）构成，氨基酸种类如表 2 所示。氨基酸混合物训练集和测试集分别如表 3 和表 4 所示，混合物中各样品比例近似相同。氨基酸纯样以及混合物的拉曼光谱均采用 SSR-200 便携式拉曼光谱仪测量得到。激光器的功率为 300 mW，线宽为 0.038 nm，波长为 784.84 nm，积分时间为 1 s，光谱范围为 0～3200 $\mathrm{cm^{-1}}$，光谱分辨率为 6 $\mathrm{cm^{-1}}$。选取 200～2000 $\mathrm{cm^{-1}}$ 范围内的光谱数据进行分析。

表 2　20 种参与谱库构建的氨基酸

Table 2　Twenty kinds of amino acids participating in the construction of spectral library

Number	Amino acid
1	Arginine
2	Proline
3	Alanine
4	Phenylalanine
5	Cysteine
6	Asparagine
7	Glutamine
8	Leucine
9	Threonine
10	Valine
11	Isoleucine
12	Glutamic acid
13	Glycine
14	Methionine
15	Lysine
16	Aspartic acid
17	Histidine
18	Serine
19	Tryptophan
20	Tyrosine

表 3　训练集中的氨基酸混合物

Table 3　Amino acid mixtures in training set

Number	Amino acid
1	Glycine and phenylalanine
2	Arginine and histidine
3	Alanine and aspartic acid
4	Glutamine and glutamic acid
5	Arginine and serine
6	Arginine, aspartic acid and proline
7	Glycine, histidine and serine
8	Glutamine, proline and leucine
9	Asparagine, glutamic acid and glycine
10	Asparagine, alanine and cysteine
11	Proline, glutamine, glutamic acid and histidine
12	Serine, arginine, alanine and leucine
13	Histidine, threonine, valine and leucine
14	Histidine, isoleucine, cysteine and lysine
15	Asparagine, glycine, aspartic acid and tryptophan

3.2　基于氨基酸混合物谱库的判别实验

3.2.1　拉曼光谱谱库的建立

采用上述拉曼光谱仪对 20 种氨基酸纯样样品进行拉曼光谱的检测。依次取少量氨基酸样品，无需对样品进行其他处理，在称量纸上轻轻压实，采集 10 次，保留平均光谱。通过广义 Whittaker 平滑器对平均光谱信号进行平滑处理[16]，并进行强度归一化。采用 VTPspline 基线背景扣除算法扣除归一化后拉曼光谱的背景基线[17]。采用 MSPD 寻峰算法对光谱信号进行寻峰[13]，使用基于 Voigt 函数拟合的拉曼谱峰判别方法对所寻谱峰进行判别[14]，进而根据(1)式得到每个氨基酸纯样拉曼光谱中每一个谱峰的权重 $\omega_{\mathrm{peak}}(i)$。

表 4　验证集中的氨基酸混合物

Table 4　Amino acid mixtures in verification set

Number	Amino acid
1	Arginine and asparagine
2	Serine and glutamic acid
3	Asparagine and glutamine
4	Arginine and alanine
5	Serine and glutamic acid
6	Arginine, histidine and aspartic acid
7	Asparagine, glutamic acid and phenylalanine
8	Asparagine, serine and glutamic acid
9	Arginine, glycine, alanine and proline
10	Arginine and asparagine, phenylalanine and glutamine
11	Aspartic acid and alanine, glutamic acid and phenylalanine

3.2.2　氨基酸混合物拉曼光谱的检测与模型建立

针对表 3 所示训练集中的氨基酸混合物依次采样 10 次，并将平均光谱作为混合物光谱。通过广义 Whittaker 平滑器对平均光谱信号进行平滑处理，之后进行强度归一化处理。采用 VTPspline 基线背景扣除算法扣除归一化后拉曼光谱的背景基线。处理后的混合物光谱与纯净物组分的光谱数据对比图如图 1 所示。

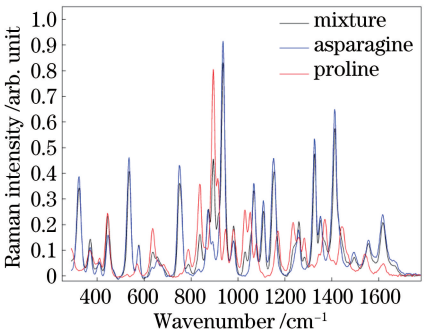

图 1　天冬酰胺、脯氨酸及其混合物的拉曼光谱

Fig. 1　Raman spectra of asparagine, proline and their mixtures

采用 MSPD 寻峰算法对光谱信号进行寻峰，使用基于 Voigt 函数拟合的拉曼谱峰判别方法对所寻谱峰进行判别。寻峰结束后，使用逆检索方法将谱库中物质的谱峰信息与被测样本谱峰信息逐一进行

比对(谱峰间距相差 6 cm^{-1} 以内即为谱峰匹配成功),进而得到谱库中可能含有的物质成分,并构成二级谱库。同时,计算得到了二级谱库中每种物质与被测样本拉曼光谱的谱峰匹配系数。根据非负最小二乘算法计算得到二级谱库中每种物质的非负最小二乘匹配系数,用以表征物质光谱对被测样品光谱的贡献程度。根据(5)式计算二级谱库中每种物质与被测样品光谱的夹角余弦匹配系数。

将依据训练集中氨基酸混合物所得的三种光谱匹配系数与实际结果组成向量 $\boldsymbol{L}_i$:

$$\boldsymbol{L}_i=[\boldsymbol{X},y_i]=[1,x_{1i},x_{2i},x_{3i},y_i], \quad (12)$$

式中:x_{1i} 为光谱谱峰的匹配系数;x_{2i} 为非负最小二乘匹配系数;x_{3i} 为夹角余弦匹配系数;$y_i\in[0,1]$,0 表示该谱库物质不存在于被测样品中,1 表示谱库物质存在于被测样品中。将所有向量组成矩阵,采用逻辑回归数学模型对矩阵数据进行处理,得到各匹配系数的权重系数 θ_i。

同时,基于谱峰匹配系数、非负最小二乘匹配系数、夹角余弦匹配系数对训练集中的氨基酸混合物光谱进行谱库匹配实验,再单独使用每种判别参数用来确定合适的阈值。使用单一匹配系数与光谱集成匹配系数的匹配效果对比如图 2 所示,圆圈表示参与匹配的谱库物质存在于被测样品中,星号表示参与匹配的谱库物质不存在于被测样品中。图中横轴为 4 种匹配系数的阈值,可认为大于某一阈值时即判断该样品存在于被检测混合物中。从图 2 中可以看出:当单独使用三种匹配系数时,均无法设置合适的阈值对谱库物质进行识别;当使用光谱集成匹配系数对谱库物质进行匹配时,两类样本是可以实现线性可分的。

训练样本集使用三种判别系数进行判别的样本分布图如图 3 所示。由于三种匹配系数之间存在一定的不相关性,因此可以从多个维度对谱库物质进行更加有效的匹配。

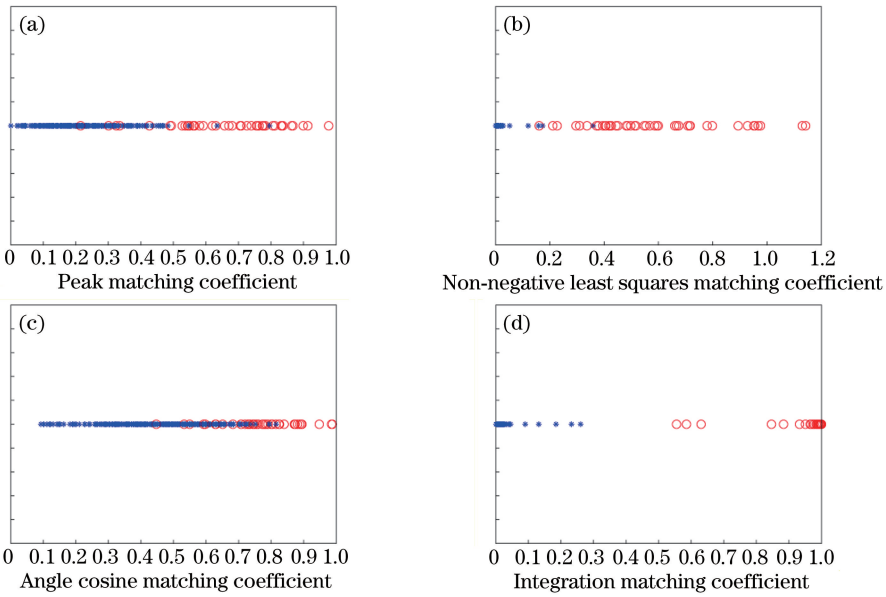

图 2　训练集样本分别使用 3 种单一匹配系数与光谱集成匹配系数的匹配效果。
(a)谱峰匹配系数;(b)非负最小二乘匹配系数;(c)夹角余弦匹配系数;(d)光谱集成匹配系数

Fig. 2　Matching results of three kinds of single matching coefficients and spectral integrated matching coefficient for samples in the training set. (a) Spectral peak matching coefficient; (b) non-negative least squares matching coefficient; (c) angle cosine matching coefficient; (d) spectral integration matching coefficient

根据训练样本的训练结果得到的逻辑回归模型参数向量为

$$\boldsymbol{\theta}=[\theta_0,\theta_1,\theta_2,\theta_3]=[-8.4836,4.2546,4.99,5.05]。 \quad (13)$$

光谱集成匹配系数 $h_\theta(\boldsymbol{X})=\dfrac{1}{1+\exp(-\boldsymbol{X}\cdot\boldsymbol{\theta})}$,若 $h_\theta(\boldsymbol{X})$ 的数值大于 0.5,则认为参与匹配的谱库物质存在于被测样品中;若 $h_\theta(X)$ 的数值小于 0.5,则认为参与匹配的谱库物质不存在于被测样品中。从图 2(d)中可以看出,使用光谱集成匹配系数可以很好地对谱库物质进行匹配和鉴别。

3.2.3　验证集拉曼光谱谱库的判别实验

针对表 4 所示验证集中的氨基酸混合物,每个样品依次采样 10 次,保留平均光谱作为样品的拉曼

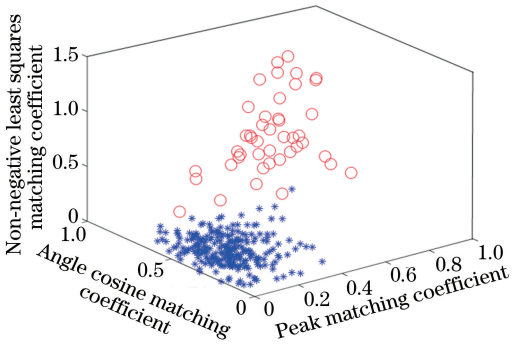

图 3　训练样本集在三种判别系数下的样本分布图

Fig. 3　Sample distribution of training samples at three discriminant coefficients

光谱。光谱处理方法与上述训练集中氨基酸混合物的光谱处理方法相同。基于谱峰匹配系数、非负最小二乘匹配系数、夹角余弦匹配系数、光谱集成匹配系数 4 种相似度测度进行谱库检索实验，并对实验结果进行对比。使用单一匹配系数与光谱集成匹配系数的匹配效果对比如图 4 所示，其中：圆圈表示参与匹配的谱库物质存在于被测样品中，星号表示参与匹配的谱库物质不存在于被测样品中。

从图 4(d)中可以看出：针对验证集样本数据，使用光谱集成匹配系数可以很好地实现被测样品与谱库物质的匹配，没有出现漏判和误判；当光谱集成匹配系数大于 0.5 时，判定该谱库物质存在于被测样品中；当匹配系数小于 0.5 时，判定该谱库物质不存在于被测样品中。使用单一匹配系数进行谱库匹配时仍然无法设定合适的阈值进行判定。

针对测试集样本数据，分别采用三种匹配数和集成系数进行谱库匹配判别。谱峰匹配系数阈值由 0.1 增加到0.9，非负最小二乘匹配系数阈值由 0.01 增加至 0.9，夹角余弦匹配系数阈值由 0.1 增加到 0.9，集成匹配系数阈值由 0.01 增加至 0.9。通过计算 4 种匹配系数在不同阈值下的 F_{DR} 与 T_{PR} 值，绘制相应的 F_{DR}-T_{PR} 曲线，结果如图 5 所示。

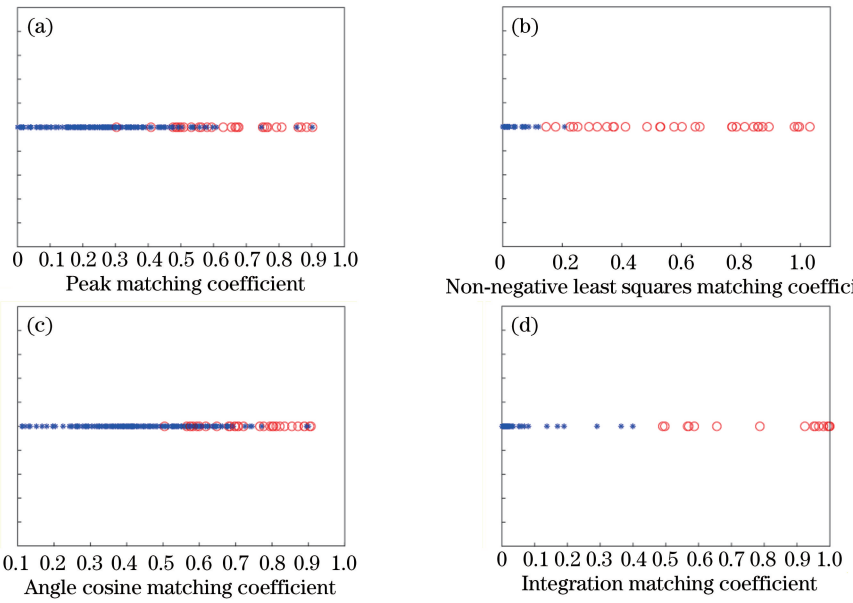

图 4　验证集样本分别使用 3 种单一匹配系数与光谱集成匹配系数的匹配效果。(a)谱峰匹配系数；(b)非负最小二乘匹配系数；(c)夹角余弦匹配系数；(d)光谱集成匹配系数

Fig. 4　Matching results of three kinds of single matching coefficients and spectral integrated matching coefficient for samples in the validation set. (a) Spectral peak matching coefficient; (b) non-negative least squares matching coefficient; (c) angle cosine matching coefficient; (d) spectral integration matching coefficient

3.3　结果与分析

在基于拉曼光谱的谱库物质识别中，混合物的准确识别一直是一大难题。由于混合物光谱谱峰重叠，且光谱更为复杂，很难采用单一匹配模型(例如谱峰匹配系数或夹角余弦匹配系数)对混合物成分进行有效识别。

从图 5 中可以看出：采用三种谱库匹配方法以及光谱集成匹配方法进行混合物识别时，虽然非负最小二乘匹配系数相较于谱峰匹配系数和夹角余弦匹配系数，在 F_{DR} 值相同的情况下拥有更高的 T_{PR} 值，但仍然存在误判的情况；而光谱集成匹配系数则可以将正负样本完全分开，没有发生误判。由于光谱集成匹配算法综合考虑了光谱谱峰信息和全谱信息，故而相较于单一匹配算法具有更高的识别率和

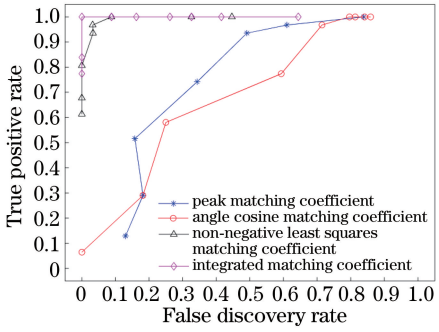

图 5　4 种匹配方法的 F_{DR}-T_{PR} 对比图

Fig. 5　F_{DR}-T_{PR} curves for four kinds of matching methods

更低的误判率，且更加稳定。

4　结　　论

为了完成基于拉曼光谱谱库的混合物组分识别，构造了一种光谱集成匹配系数。该匹配系数通过逻辑回归数学模型融合了谱峰匹配系数、非负最小二乘匹配系数、夹角余弦匹配系数三种匹配模型，通过对样本数据集进行训练，得到了相应的权重系数。使用该权重系数可以计算得到光谱的集成匹配系数，用以判别被测样本中是否含有该组分。集成光谱匹配系数融合了光谱特征谱峰信息与全谱信息。采用氨基酸混合物拉曼光谱进行验证实验，结果表明，集成拉曼光谱匹配系数可以更加有效地判别混合物的组分。希望该方法能对基于谱库的混合物光谱匹配研究提供思路和参考。

参　考　文　献

[1] Tamor M A, Vassell W C. Raman "fingerprinting" of amorphous carbon films[J]. Journal of Applied Physics, 1994, 76(6): 3823-3830.

[2] Wang S, Zeng H S. Real-time *in vivo* Raman spectroscopy and its clinical application in early cancer detection[J.] Chinese Journal of Lasers, 2018, 45(2): 0207002.
王爽, Zeng Haishan. 实时拉曼光谱分析技术及其在临床早期癌症检测中的应用[J]. 中国激光, 2018, 45(2): 0207002.

[3] Fang X Q, Peng Y K, Li Y Y, *et al*. Rapid and quantitative detection method of sodium benzoate in carbonated beverage based on surface-enhanced Raman spectroscopy[J]. Acta Optica Sinica, 2017, 37(9): 0930001.
房晓倩, 彭彦昆, 李永玉, 等. 基于表面增强拉曼光谱快速定量检测碳酸饮料中苯甲酸钠的方法[J]. 光学学报, 2017, 37(9): 0930001.

[4] Blanco M, Romero M A. Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation[J]. The Analyst, 2001, 126(12): 2212-2217.

[5] Chen Z P, Li L M, Jin J W, *et al*. Quantitative analysis of powder mixtures by Raman spectrometry: the influence of particle size and its correction[J]. Analytical Chemistry, 2012, 84(9): 4088-4094.

[6] Sato-Berrú R Y, Medina-Valtierra J, Medina-Gutiérrez C, *et al*. Quantitative NIR-Raman analysis of methyl-parathion pesticide microdroplets on aluminium substrates[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2004, 60(10): 2231-2234.

[7] Zhou W H, Ying Y B, Xie L J. Spectral database systems: a review[J]. Applied Spectroscopy Reviews, 2012, 47(8): 654-670.

[8] Zhang Z M, Chen X Q, Lu H M, *et al*. Mixture analysis using reverse searching and non-negative least squares[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 137: 10-20.

[9] Gawinkowski S, Kamińska A, Roliński T, *et al*. A new algorithm for identification of components in a mixture: application to Raman spectra of solid amino acids[J]. The Analyst, 2014, 139(22): 5755-5764.

[10] Peng Y, Zhang Z M, Lu H M, *et al*. Identification of colorants in food by surface-enhanced Raman spectroscopy and wavelet-based reverse search[J]. Journal of Instrumental Analysis, 2017, 36(5): 627-632.
彭颖, 张志敏, 卢红梅, 等. 基于小波-反向搜索及表面增强拉曼的食品中色素的光谱定性分析[J]. 分析测试学报, 2017, 36(5): 627-632.

[11] Kong X B, Shu N, Tao J B, *et al*. A new spectral similarity measure based on multiple features integration[J]. Spectroscopy and Spectral Analysis, 2011, 31(8): 2166-2170.
孔祥兵, 舒宁, 陶建斌, 等. 一种基于多特征融合的新型光谱相似性测度[J]. 光谱学与光谱分析, 2011, 31(8): 2166-2170.

[12] Chu X L, Li J Y, Chen P, *et al*. Algorithms, strategies and application progress of spectral searching methods[J]. Chinese Journal of Analytical Chemistry, 2014, 42(9): 1379-1386.
褚小立, 李敬岩, 陈瀑, 等. 分子光谱自动检索算法、策略与应用进展[J]. 分析化学, 2014, 42(9): 1379-1386.

[13] Zhang Z M, Tong X, Peng Y, *et al*. Multiscale peak detection in wavelet space[J]. The Analyst, 2015, 140(23): 7955-7964.

[14] Liu M H, Dong Z R, Xin G F, *et al*. Discrimination method of Raman spectral peaks based on Voigt

function fitting[J]. Chinese Journal of Lasers, 2017, 44(5): 0511003.
刘铭晖, 董作人, 辛国锋, 等. 基于 Voigt 函数拟合的拉曼光谱谱峰判别方法[J]. 中国激光, 2017, 44(5): 0511003.

[15] Lawson C L, Hanson R J. Solving least squares problems[M]. [S. l]: Society for Industrial and Applied Mathematics, 1995: 160-165.

[16] Eilers P H C. A perfect smoother[J]. Analytical Chemistry, 2003, 75(14): 3631-3636.

[17] Cai Y Y, Yang C H, Xu D G, *et al*. Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation [J]. Analytical Methods, 2018, 10(28): 3525-3533.