中红外光谱法结合支持向量机快速鉴别蜂蜜品种

徐天扬[1,2,3]，杨娟[1]，孙晓荣[4,5]，刘翠玲[4,5]，李熠[1,2,3]，周金慧[1,2,3]，陈兰珍[1,2,3]*

[1]中国农业科学院蜜蜂研究所，北京 100093；

[2]农业部蜂产品质量安全控制重点实验室(北京)，北京 100093；

[3]农业部蜂产品质量安全风险评估实验室，北京 100093；

[4]北京工商大学计算机与信息工程学院，北京 100048；

[5]食品安全大数据技术北京市重点实验室，北京 100048

摘要　为快速鉴别5种蜂蜜(椴树蜜、荆条蜜、油菜蜜、洋槐蜜、荔枝蜜)的品种，首次提出了基于主成分分析(PCA)方法结合线性支持向量机(SVM)或最小二乘支持向量机(LSSVM)的中红外光谱法鉴别蜂蜜品种的新方法。用傅里叶变换中红外光谱仪测定5种蜂蜜样本的中红外光谱，并进行归一化预处理，然后用主成分分析降维方法分别提取经预处理后的光谱数据中的5维、10维、15维、20维特征数据，最后设计了线性SVM和基于网格搜索优化算法的径向基函数(RBF)的LSSVM分类器模型。利用不同分类器模型，识别未知蜂蜜样本光谱数据降维到不同维数的特征数据，并进行实验验证。结果表明：应用主成分分析降维方法降维到20维的特征数据在SVM和LSSVM分类器上的平均识别率均高于97%，最高识别率均可达到100%，且稳定性很好；利用较低维数数据进行分类时，LSSVM分类器比SVM的识别精度更高，稳定性更好。研究证明将中红外光谱与线性SVM或LSSVM结合用于快速鉴别蜂蜜品种是可行的。

关键词　光谱学；中红外光谱；主成分分析；支持向量机；最小二乘支持向量机；径向基函数

中图分类号　O657.33　　**文献标识码**　A　　**doi**：10.3788/LOP55.063003

Mid-Infrared Spectroscopy Analysis Combined with Support Vector Machine for Rapid Discrimination of Botanical Origin of Honey

Xu Tianyang[1,2,3], Yang Juan[1], Sun Xiaorong[4,5], Liu Cuiling[4,5], Li Yi[1,2,3], Zhou Jinhui[1,2,3], Chen Lanzhen[1,2,3]

[1] *Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing* 100093, *China*;

[2] *Key Laboratory of Bee Products for Quality and Safety Control, Ministry of Agriculture, Beijing* 100093, *China*;

[3] *Laboratory of Risk Assessment for Quality and Safety of Bee Products, Ministry of Agriculture, Beijing* 100093, *China*;

[4] *School of Computer and Information Engineer, Beijing Technology and Business University, Beijing* 100048, *China*;

[5] *Beijing Key Laboratory of Large Data Technology for Food Safety, Beijing* 100048, *China*

Abstract　To achieve the fast discrimination of five varieties of honeys, namely linden honey, vitex honey, rape honey, acacia honey and litchi honey, we propose a new method in this article by using the mid-infrared spectra based on principle component analysis (PCA) combined with linear support vector machine (SVM) or least squares support vector machine (LSSVM). The mid-infrared spectra of five varieties of honey samples are determined by Fourier transform infrared spectroscopy and normalized. Then the 5-dimensional, 10-dimensional, 15-dimensional, and 20-dimensional feature data will be extracted from spectra with the use of dimension reduction method of PCA after normalization. Finally, the two classifier models, linear SVM and LSSVM with radial basis function (RBF)

收稿日期：2017-11-23；**收到修改稿日期**：2018-01-05

基金项目：国家自然科学基金面上项目(331772070)、中国农业科学院创新工程项目(CAAS-ASTIP-2017-IAR)、国家特色农产品风险评估专项(GJFP2017010)、国家蜂产业技术体系(CARS-45-KXJ10)

作者简介：徐天扬(1985—)，女，硕士，科研助理，主要从事数据处理与模式识别方面的研究。E-mail：goodgoodstudy0929@126.com

* **通信联系人**。E-mail：chenlanzhen2005@126.com

based on the grid search optimization, are designed. Using different classifier model, we identify the different dimensional feature data extracted from spectra data of unknown honey samples. Then the results of different dimension feature data and different support vector machines are validated. Experimental results show that for the 20-dimensional feature data obtained by the dimension reduction method of PCA, an average recognition rate of higher than 97% on SVM and LSSVM classifiers is achieved, the highest recognition rate can reach 100%, and classifier stability is very good. LSSVM classifier has higher recognition accuracy and better stability than linear SVM classifier in classification with lower dimension data. Hence, it proves the feasibility of rapid identification of five varieties of honeys with mid-infrared spectra combined with linear SVM or LSSVM.

Key words spectroscopy; mid-infrared spectrum; principle component analysis; support vector machine (SVM); least squares support vector machine (LSSVM); radial basis function (RBF)

OCIS codes 300.6170; 300.6340; 200.4560

1 引 言

蜂蜜作为日常生活中最常饮用的、在国内外市场上销售越来越普及的一种营养品，如今已倍受消费者的欢迎。由于蜜蜂所采植物花蜜不同，所以不同蜂蜜具有的食用价值、保健功效和医用功效也不同。目前，在与蜂蜜分类方法相关的国际标准中，最为典型的就是由国际食品法典委员会和欧盟制定的相关分类指令。分析这些分类指令可以发现，这两大机构在蜂蜜划分标准上还是以蜜源来自的植物作为分类依据的，由此可以根据蜜源和产地对蜂蜜进行划分，这也是蜂蜜最常见的分类方法。不同蜜源的蜂蜜不仅在内在品质和感官上具有明显差异，其所含有的营养价值也不同，最终导致价格也有很大差别[1-4]。针对目前的这种情况，如果能够快速将不同种类和品级的蜂蜜区分开来，将有助于蜂蜜产品实现快速市场定价，并指导消费者买到货真价实的蜂蜜产品，从而促进整个蜂蜜行业市场秩序的完善。

感官评定是鉴定蜂蜜最传统、应用时间最长的方法，具有简单方便、鉴定过程中耗用的机器成本低等优点，但却需要专业性强、经验丰富的评价人员。同时，蜂蜜的色、香、味等感官特征很容易受到外界环境和人为因素的干扰，如采集方式、储存方式、结晶条件等的改变都会降低评定结果的客观准确性。另一种鉴别蜂蜜品种的传统方法是花粉显微镜法，就是通过显微镜鉴别蜂蜜中花粉的数量和类型来确定蜂蜜品种。该方法成本高、耗时长，且受多种因素的影响，检测的准确度较低。其中主要的影响因素有花粉原产地、蜂蜜的质量和采集方式、专家自身的经验等[2]。研究人员开发出了利用客观分析技术(如质谱分析法和色谱分析法等)对蜂蜜进行鉴别的检测方法，虽然这些检测技术能够准确、有依据地判断蜂蜜所属的品种，但是检测程序较为繁琐，成本高，耗时长[3]。于是，快速、低成本的光谱鉴别蜂蜜品种技术应势而生，这些技术主要有近红外、中红外、拉曼、核磁共振光谱技术。与质谱、色谱技术等相比，光谱技术前处理更简单、环保，而且能获得更为丰富的样本信息。

中红外光谱与近红外光谱都能显现有机物的分子振动信息，但中红外光谱的检测限比近红外要高1～2个数量级，同时还具有吸收峰窄、谱峰重叠不严重、信息量较大、信息提取更容易、样品信息表达更丰富、分子选择性更好等优点[5]。在农产品品种鉴别和产地溯源方面，该技术结合化学计量学的方法已得到了广泛应用，并取得较大进展[6]：文献[7]利用中红外光谱技术结合线性判别分析方法建立了6种蜜源蜂蜜的数学鉴别模型；文献[8]应用中红外分析仪器和模式识别技术针对掺入糖浆的洋槐蜂蜜和紫云英蜂蜜与纯蜂蜜的品质差异建立了模式识别分类研究模型；文献[9]通过对饶河本地和其他地区蜂蜜样本的中红外光谱谱图进行分析，利用化学计量软件建立了饶河黑蜂蜂蜜产地真假判别模型，该模型的判定率达到了90.3%。

本文应用中红外光谱技术结合统计学分析中经典的主成分分析降维方法，针对椴树蜜、荆条蜜、油菜蜜、洋槐蜜、荔枝蜜5种蜂蜜构建线性支持向量机(SVM)和最小二乘支持向量机(LSSVM)分类判别模型；应用这2种模型对130个测试样本进行鉴别，其平均识别率均高于97%，最高识别率均可达到100%，取得了较为理想的分类效果。

2 基本原理

2.1 降维方法原理

数据降维一方面可以解决“维数灾难”，降低复杂度；另一方面可以减少冗余信息造成的误差，提高识别精度。主成分分析(PCA)方法是常用的一种降维方法，其原理是运用线性映射将 n 维特征映射到 k 维上($k<n$)，这 k 维全新的正交特征被称为主元，

在无损或很少损失数据集信息量的情况下，降低了数据集的维数，较好地保留了原数据集的主要信息，是一种常用的将一组可能存在相关性的变量转换为一组线性不相关的变量的方法，被广泛应用于各领域[10-12]。文献[13-14]将主成分分析降维方法应用于光谱数据处理领域。由于天然形成的蜂蜜具有十分复杂的组成成分，因此蜂蜜中红外光谱的形成过程中存在着各种成分吸收光度叠加的情况，因此共线性是一定存在的，在分析中如果不能合理地处理这些共线性的情况，最终的分析结果将会出现偏差。处理共线性的常用方法就是在建模前先进行主成分分析。通过主成分分析不仅能够基本去除光谱矩阵中的共线性关系和无用的干扰信息，还能对光谱矩阵进行降维，精简优化模型。本研究就是采用主成分分析方法对标准归一化后的原始数据进行降维处理的。

2.2 SVM 原理

SVM 方法是一种较为成熟的分类方法，这种分析方法的主要优点是能够解决模式识别中的小样本、非线性分类问题，还能够有效处理高维模型。SVM 的学习问题最终是以凸优化问题的表示形式出现的，从而可以利用已知的有效算法计算出目标函数的全局最小值，以解决最优解问题。这一全局最优解的获得是一些其他算法无法得到的。以基于规则的分类和神经网络模型为例，它们都是基于贪心学习的策略，只能通过计算得到局部最优解，而不是全局最优解。因此本研究选用 SVM 方法进行模式识别，以获得全局最优解。

总体来讲，SVM 原理[15]是以寻找一个最优超平面为目的的，最优的意思即两类样本能够尽可能地被划分到超平面两侧，而且使超平面到超平面两侧的数据点的距离最大。解决这个问题时，将求解最优分类面的最优化问题转化为其对偶问题，从而通过求解相对简单的对偶问题来实现求解原分类问题的算法。本研究选用线性 SVM 方法和非线性径向基函数(RBF)核 LSSVM 方法进行对比，以验证不同 SVM 的效果。

线性 SVM 方法原理[16]如下所述。在样本空间中，假设线性可分样本集为(x_i, y_i)，其中，样本数 $i=1,\cdots,l$；类别标号为 $y=\{+1,-1\}$，划分超平面的表达式为

$$g(x)=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b=0, \tag{1}$$

式中：$g(x)$为分类函数；$\boldsymbol{\omega}$ 为法向量；$\boldsymbol{x}$ 为测试样本；b 为截距。

假设它已经完成了对样本的分隔，且两种样本的标签分别是$\{+1,-1\}$，那么对于一个分类器来说，$g(x)>0$ 和 $g(x)<0$ 就可以分别代表两个不同的类别：+1 和−1。

为尽最大努力使分开的两个类别有最大的间隔，并使分隔具有更高的可信度，以及对未知的新样本有很好的分类预测能力(在机器学习中被称为泛化能力)，需要使离分隔面最近的数据点具有最大的距离。为了描述离分隔超平面最近的数据点，需要找到两个和这个超平面平行和距离相等的超平面：H_1 和 H_2，即

$$y_{H_1}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b=+1, \tag{2}$$

$$y_{H_2}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b=-1。 \tag{3}$$

在这两个超平面上的样本点也就是理论上离分隔超平面最近的点，它们的存在决定了 H_1 和 H_2 的位置，支撑起了分界线，这些样本点就是所谓的支持向量。由(2)～(3)式可以推出两个超平面(H_1 和 H_2)的间隔为 $2/\|\boldsymbol{\omega}\|$，即现在的目的是实现这个间隔最大化，相当于最小化 $\|\boldsymbol{\omega}\|$，为了之后的求导和计算方便，相当于进一步最小化 $\|\boldsymbol{\omega}\|^2/2$。

假设超平面能将样本正确分类，则可令

$$\begin{cases}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i+b\geqslant+1, & y_i=+1\\ \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i+b\leqslant-1, & y_i=-1\end{cases}。 \tag{4}$$

式中：$\boldsymbol{x}_i$ 为训练样本。

(4)式和(5)式合并后可以得到：

$$y_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1, \tag{5}$$

这就是目标函数的约束条件。现在这个问题就变成了一个最优化问题：

$$\min_{\boldsymbol{\omega},b}\frac{1}{2}\|\boldsymbol{\omega}\|^2, \tag{6}$$

$$y_i[(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i)+b]-1\geqslant 0(i=1,2,\cdots,l)。 \tag{7}$$

以上是典型的二次凸规划问题。根据目标函数的特点和给出的约束条件，可知其具有凸性，可以利用最优化理论获得全局最小最优解，因此，(8)式和(9)式可以采用 Lagrange 乘子法，且满足 KKT 条件。(8)式和(9)式的表达式为

$$a_i[y_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i+b)-1]=0, \tag{8}$$

$$f(x)=\mathrm{sgn}\{\boldsymbol{\omega}^{*\mathrm{T}}\cdot\boldsymbol{x}+b^*\}=$$

$$\mathrm{sgn}\Big\{\sum_{i=1}^{l}a_i^*y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b^*\Big\}, \tag{9}$$

式中：a 为拉格朗日乘子；a^*、b^*、$\boldsymbol{\omega}^*$ 分别为对应于 a、b、ω 的最优解；$(\boldsymbol{x}_i\cdot\boldsymbol{x})$为 2 个向量的内积。分类函数如(9)式所示，$\boldsymbol{x}$ 的归属是根据其符号来确定的。

在现实情况中，几乎不可能存在线性可分的情形，针对某些误分点引入一种度量 ζ_i，用来描述训

练集被错划的程度。于是目标变成间隔 $2/\|\boldsymbol{\omega}\|$，其值越大，错划的程度越小，故引入惩罚函数 C 作为综合两个目标的权重，则原始问题变为

$$\min_{\boldsymbol{\omega},b}\frac{1}{2}\|\boldsymbol{\omega}^2\|+C\sum_{i=1}^{l}\zeta_i, \tag{10}$$

$$y_i[(\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{x}_i)+b]\geqslant 1-\zeta_i, i=1,\cdots,l。 \tag{11}$$

相应的 Lagrange 函数为

$$L(\boldsymbol{\omega},b,\zeta,a,r)=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{l}\zeta_i-\sum_{i=1}^{l}a_i\{y_i[(\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{x}_i)+b]-1+\zeta_i\}-\sum_{i=1}^{l}\gamma_i\zeta_i。 \tag{12}$$

对 L 关于 $\boldsymbol{\omega}$、b 和 ζ 求极小，得到

$$\sum_{i=1}^{l}a_iy_i=0, \tag{13}$$

$$\boldsymbol{\omega}=\sum_{i=1}^{l}a_iy_ix_i, \tag{14}$$

$$C-a_i-\gamma_i=0。 \tag{15}$$

然后将上述条件回代，对 a 求极大则得原问题的对偶问题：$\min_a\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_ja_ia_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{j=1}^{l}a_j$，$\sum_{i=1}^{l}y_ia_i=0$，$0\leqslant a_i\leqslant C$，$i=1,\cdots,l$，得最优解 $\boldsymbol{a}^*=(a_1^*,\cdots,a_l^*)^{\mathrm{T}}$，计算得到 $\boldsymbol{\omega}^*=\sum_{i=1}^{l}y_ia_i^*\boldsymbol{x}_i$。

选择 a^* 的一个小于 C 的正分量，并根据此计算 $b^*=y_j-\sum_{i=1}^{l}y_ia_i^*(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)$。构造超平面 $(\boldsymbol{\omega}^{*\mathrm{T}}\cdot\boldsymbol{x})+b^*=0$，求得决策函数

$$f(x)=\mathrm{sgn}\{\boldsymbol{\omega}^{*\mathrm{T}}\cdot\boldsymbol{x}+b^*\}=\mathrm{sgn}\Big\{\sum_{i=1}^{l}a_i^*y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b^*\Big\}。 \tag{16}$$

最终的分类函数如(16)式所示。这就是 Libsvm 工具箱中 C-SVC 程序采用的线性核函数，即本研究采用的第一种线性 SVM 程序的公式原理及推导。

非线性 LSSVM 方法原理[17]和线性 SVM 方法原理的区别在于 LSSVM 将原方法的不等式约束变为等式约束，从而大大方便了 Lagrange 乘子 a 的求解，原问题是二次规划(QP)问题，而在 LSSVM 中则是一个求解线性方程组的问题。非线性 LSSVM 利用映射方法，在计算开始之前先选择非线性映射，随后输入的向量将通过选择的映射关系被映射到高维特征空间中。运用最小化结构风险原则在这个空间中构建最优决策函数，并巧妙地运用原空间的核函数取代高维特征空间中的点积运算。

对于 LSSVM，原问题的不等式约束变成等式约束

$$\min_{\boldsymbol{\omega},b,e}J(\boldsymbol{\omega},e)=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+\frac{1}{2}r\sum_{i=1}^{l}e_i^2, \tag{17}$$

$$y[\boldsymbol{\omega}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b]=1-e_i, \tag{18}$$

式中：e_i 为误差控制函数。LSSVM 优化目标的损失函数是误差 e_i 的二次项。在 LSSVM 中，r 是一个权重，与 C 一样，用于平衡寻找最优超平面和偏差量的最小值。为了容许有一定的错分率，将最大分类间隔和最小错分样本折中考虑。

接下来，与 SVM 类似，采用 Lagrange 乘子法把原问题转化为对单一参数，也就是求 a 的极大值问题。新问题如下：

$$L(\boldsymbol{\omega},b,e_i,a)=J(\boldsymbol{\omega},e)-\sum_{i=1}^{l}a_i\{y_i[\boldsymbol{\omega}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b]-1+e_i\}。 \tag{19}$$

分别对 $\boldsymbol{\omega}$、b、e_i、a_i 求导，则有

$$\frac{\partial L}{\partial\boldsymbol{\omega}}=0\rightarrow\boldsymbol{\omega}=\sum_{i=1}^{l}a_iy_i\phi(\boldsymbol{x}_i), \tag{20}$$

$$\frac{\partial L}{\partial b}=0\rightarrow\sum_{i=1}^{l}a_iy_i=0, \tag{21}$$

$$\frac{\partial L}{\partial e_i}=0\rightarrow a_i=re_i, \tag{22}$$

$$\frac{\partial L}{\partial a_i}=0\rightarrow y_i[\boldsymbol{\omega}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b]-1+e_i=0。 \tag{23}$$

根据(19)式和求导后的 4 个条件可以消去 e_i 和 $\boldsymbol{\omega}$，列出一个关于 $\boldsymbol{\alpha}$ 和 b 的线性方程组

$$\begin{bmatrix}0 & \boldsymbol{y}^{\mathrm{T}}\\ \boldsymbol{y} & \boldsymbol{\Omega}_{i,j}+r^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{I}_l\end{bmatrix}, \tag{24}$$

式中：$\boldsymbol{\Omega}_{ij}$ 为核矩阵，$\boldsymbol{\Omega}_{ij}=y_iy_j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)=y_iy_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$，$j=1,\cdots,l$；$\boldsymbol{I}$ 为单位矩阵；$\boldsymbol{I}_l=[1\quad\cdots\quad l]^{\mathrm{T}}$，$\boldsymbol{\alpha}=[a_1\quad\cdots\quad a_l]^{\mathrm{T}}$。

解上述方程组可以得到一组最优分类面的参数 $\boldsymbol{\alpha}$ 和 b，最后得到 LSSVM 的分类函数为

$$y(x)=\mathrm{sgn}\Big\{\sum_{i=1}^{l}a_iy_iK(\boldsymbol{x},\boldsymbol{x}_i)+b\Big\}, \tag{25}$$

式中：$K(\boldsymbol{x},\boldsymbol{x}_i)$为核函数。

LSSVM 的核函数必须是正定的，且满足 Mercer 定理[18]。这就是 LSSVM 工具箱，即本研究采用的第二种 LSSVM 的原理以及公式推导。

一对一 SVM 多分类原理[19]是在任意两类样本之间设计 1 个 SVM，为任意两类构建超平面，因此

针对 n 个类别的样本，就需要设计 $n(n-1)/2$ 个 SVM。在这种方式下，是对 n 个分类器的训练集进行两两区分。测试时，当要判别一个未知样本的类别时，需要进行投票，最后投票计数最多的类别即判定为该未知样本的类别。按如下方式进行投票（A、B、C、D 分别为未知样本的投票类别基数）：

$A=B=C=D=0$；

(A,B)-SVM，如果是 A 获胜，则 $A=A+1$；否则，$B=B+1$；

(A,C)-SVM，如果是 A 获胜，则 $A=A+1$；否则，$C=C+1$；

…

(C,D)-SVM，如果是 C 获胜，则 $C=C+1$；否则，$D=D+1$；

最大(A,B,C,D)就是未知样本的决策类别。

本研究采用 Libsvm 工具箱中的 C-SVC（线性核）方法和 LSSVM 工具箱中的 LSSVM（RBF 核）方法中的一对一机制对 5 类蜂蜜样本进行多分类，并验证分类效果。

3 材料与方法

3.1 材　　料

采用来自不同蜂场的蜂蜜样本，共 392 个，其中椴树蜜 39 个、荆条蜜 47 个、油菜蜜 74 个、洋槐蜜 86 个、荔枝蜜 146 个。每个样品均在 20 ℃左右密封保存。

3.2 仪器设备

应用德国布鲁克公司生产的配备衰减全反射 ATR 附件的 TENSOR37 傅里叶变换中红外光谱仪，及该公司的光谱采集和分析软件 OPUS7.0。

3.3 光谱采集

将少量蜂蜜样本涂于 ATR 附件上，取每个样本扫描两次的平均值。仪器的检测参数如下：扫描范围为650～4000 cm^{-1}，分辨率为8 cm^{-1}，扫描 32 次。首先在 40～60 ℃下水浴处理结晶样品，待其融化为液体后再进行扫描。原始光谱如图 1 所示。

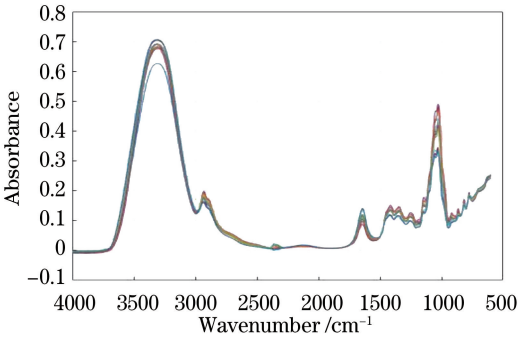

图 1 不同蜂蜜样本的中红外光谱

Fig. 1 Mid-infrared spectra of different honey samples

3.4 预处理及模型样本

影响样品光谱的因素有很多，如基线漂移、高频随机噪声、光散射等，为消除这些影响，获取有效信息，在建立判别模型之前，需要对原始光谱进行预处理。本研究采用标准归一化法对光谱进行预处理。

在模型建立时，随机选择 392 个样品中的 2/3 样本作为训练集，共 261 个样品，其余的作为测试集，共 131 个样品。椴树蜜、荆条蜜、油菜蜜、洋槐蜜、荔枝蜜依次标记为 5 到 1。

3.5 数据分析环境

用 PCA 方法将归一化后的光谱数据降到不同维数数据集之后的 4 组数据进行归一化处理，之后分别在 MATLAB2014a 和 MATLAB2016b 平台上应用线性 SVM 和非线性 LSSVM 进行分类比较。同时采用 libsvm-mat-2.89-3 和 LSSVMlabv1_8_R2009b_R2011a 软件包，该软件包具有操作简单、使用方便、通用性好的特点[20]。

4 实验结果与分析

4.1 实验结果

蜂蜜品种判别模型是根据 SVM 的判别结果来鉴别蜂蜜种类的。首先比较了不同主成分累积方差贡献率，如表 1 所示。

表 1 不同主成分累积方差贡献率

Table 1 Cumulative variance contribution rate of different principal components

Dimension	1	2	3	5	10	15	20
Contribution rate /%	46.539	82.272	91.908	96.054	99.188	99.129	99.856

由于光谱中存在着严重的共线性现象，且前 5 个主成分的累积方差贡献率为 96.054%（如表 1 所示），这一数据表明光谱中存在严重的共线性现象，因此采用主成分分析方法进行主成分分析是十分必要的。后面应用到的降维后的 n 维特征数据就是应用前 n 个主成分得分形成的矩阵。

随后对未知蜂蜜光谱样本降维到 5 维、10 维、15 维、20 维的特征数据再次进行归一化处理，应用线性 SVM 和基于网格搜索优化算法的径向基核的 LSSVM 分类器模型进行识别，验证采用不同 SVM

的效果。结果显示,在应用线性 SVM 分类器对降维到 20 维的光谱数据进行识别时,平均识别率大于 97%,最高识别率为 100%。应用 LSSVM 分类器时,需要根据特定数据通过实验的方法确定其结构和参数。故本文选取作为核函数的径向基函数 RBF、多项式函数和 Sigmoid 函数,最终发现 RBF 作为核函数时具有较高的分类精度,从而选取 RBF 作为核函数。

径向基函数 RBF 表示为

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_i|^2}{2\delta^2}\right), \tag{26}$$

$$|\boldsymbol{x}-\boldsymbol{x}_i|=\sqrt{\sum_{k=1}^{n}(\boldsymbol{x}^k-\boldsymbol{x}_i^k)^2}, \tag{27}$$

式中:δ 为核宽度。

r 和 δ 是 RBF 核函数涉及的两个主要参数。用于控制错分样本惩罚程度的惩罚参数 r 可以起到保持样本偏差与机器泛化能力之间平衡的作用。另一个重要参数 δ 是径向基核宽度,当其值过小时,样本数据会产生过学习现象;当其过大时,则会产生欠学习的现象[16]。本研究从参数集中选取 r 和 δ 的不同参数组合,采用网格搜索寻优法在全局寻找最优解[21]。通过该方法训练 LSSVM,若参数组合对应最高识别率,则为两个参数的最优组合。结果发现,降维到 20 维的特征数据在 LSSVM 分类器上的平均识别率为 97%,且在 $r=0.9538$、$\delta=22.5241$ 的条件下最高识别率达到了 100%。结果如表 2 所示。

表 2 线性 SVM 和 LSSVM 分类器模型对不同维数特征数据的平均识别率

Table 2 Average discrimination rate of different dimension feature data from linear SVM and LSSVM classifier models %

Method	5-dimension	10-dimension	15-dimension	20-dimension
PCA-SVM	84.54	87.38	96.38	97.77
PCA-LSSVM	87.31	92.54	96.15	97.69

线性 SVM 和 LSSVM 分类器模型测试集的应用结果如图 2~5 所示。

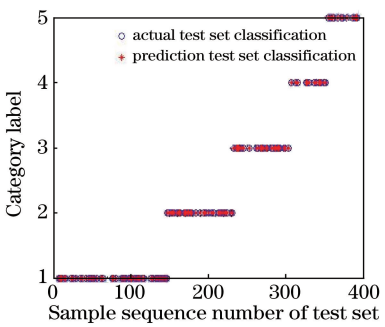

图 2 当输入 20 维特征数据且应用 SVM 算法识别率为 100%时测试集的实际分类和预测分类结果

Fig. 2 Actual and predicted classifications of test set using SVM algorithm when recognition rate is 100% and 20-dimensional feature data are input

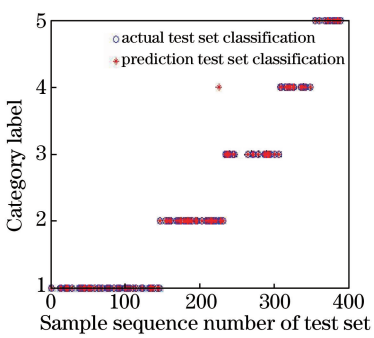

图 3 当输入为 20 维特征数据且应用 SVM 算法识别率为 99.23%时测试集的实际分类和预测分类结果

Fig. 3 Actual and predicted classifications of test set using SVM algorithm when recognition rate is 99.23% and 20-dimensional feature data are input

由图 2 和图 4 可知,椴树蜜、荆条蜜、油菜蜜、洋槐蜜、荔枝蜜的实际测试集分类和预测测试集分类完全相同,因此每种蜜的识别率都为 100%。图 3 中只有第二类洋槐蜜的 1 个样本的预测种类与实际种类不符,这个洋槐蜜样本被误判为第四类荆条蜜。图 5 中只有第二类洋槐蜜的 2 个样本的预测种类与实际种类不符,其中,1 个洋槐蜜样本被误判为第四类荆条蜜,另外 1 个洋槐蜜样本被误判为第一类荔枝蜜。第五类椴树蜜的 1 个样本的预测种类与实际种类不符,被误判为第四类荆条蜜。

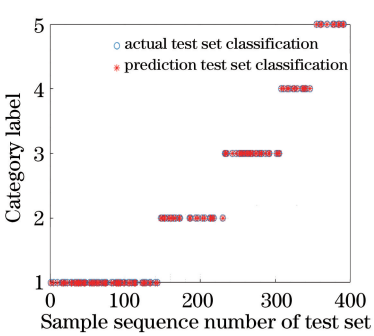

图 4 当输入为 20 维特征数据且应用 LSSVM 算法识别率为 100%时测试试集的实际分类和预测分类结果

Fig. 4 Actual and predicted classifications of test set using LSSVM algorithm when recognition rate is 100% and 20-dimensional feature data are input

4.2 结果分析

对于 PCA 降维方法,累积方差贡献率随着降维

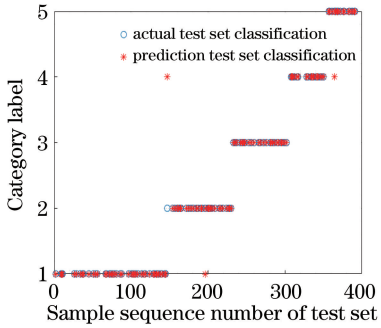

图 5 当输入为 20 维特征数据且应用 LSSVM 算法识别率为 97.69%时测试集的实际分类和预测分类结果

Fig. 5 Actual and prediction classifications of test set using LSSVM algorithm when recognition rate is 97.69% and 20-dimensional feature data are input

维数的增大而增大，这是因为降维维数增大意味着选择了更多的主成分来表示原数据集的主要信息，进而特征提取率逐渐增大。

应用较高维数降维数据结合线性 SVM 方法进行分类的效果好于 LSSVM 方法。高维(15 维、20 维)矩阵输入线性 SVM 和 LSSVM 进行分类的效果差不多，都能达到高于 96%的识别率，输入 20 维矩阵的识别率最高能达到 100%。主要原因可能是原始数据经主成分分析方法降维后在高维(15 维、20 维)矩阵上的特征提取得好，包含的信息量足够大，包含差异性信息的重要样本点在其数据空间上可以近似认为是线性可分的。这时使用线性核函数 SVM 可以达到较为理想的分类效果。而低维(5 维、10 维)矩阵输入线性 SVM 进行分类的效果与 LSSVM 的效果有一定差距。因为低维(5 维、10 维)矩阵的信息量相对较少，包含差异性信息的重要样本点在其数据空间上并不一定是线性可分的，这在一定程度上影响了分类效果。采用 RBF 核的 LSSVM 的主要计算方法就是通过非线性映射使输入的向量从原空间映射到高维空间，而在这个高维空间中得到的线性判别函数可以代替原来空间中非线性的此类函数，即在这个新的空间中，可以比原空间更容易获得最优线性分类面，这就是 RBF 核函数的贡献。Vapnik 和 Shapire 证明了假定数据本身无噪声时，理论上会存在一个核函数，使映射到高维数据空间后的数据线性可分，且高维投影通过增加空间灵活度和减少支持向量提高了测试集上低错误率的保证。依据 SVM 的性质可以获得如下的误差期望上界：

$$\{L_{\text{test}}\} \leqslant \frac{N_{\text{SV}}}{N}, \tag{28}$$

式中：$\{L_{\text{test}}\}$为测试集上错误率的期望；N_{SV}为支持向量数目；N 为样本总数。(28)式表明，支持向量越多，测试集的错误率可能会越高。这表明：如果数据离分界面很远，则支持向量一般会很少，测试性能应当可靠；反之，如果数据全部分布在分界面上，表明数据缠绕很严重，测试集基本不可靠。

在线性不可分的情况下，一切线性不可分的样本点，也就是 $\zeta_i \geqslant 0$ 的样本点，都可以作为支持向量。在 1 维的情况下，所有交叠的样本点全部都是支持向量，这样的测试集的错误率基本没有保证；反之，如果把这些交叠的样本点投射到 2 维空间，那么它们中只有一部分是支持向量，而其余交叠的样本点由于空间自由度提高，ζ_i 会发生戏剧性的变化，如图 6 所示。

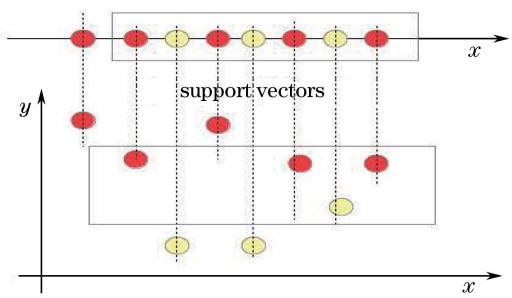

图 6 支持向量由 1 维投影到 2 维空间的变换图

Fig. 6 Transformation graph of support vector from 1-dimensional space to 2-dimensional space

1 维中交叠的那些点被投影到 2 维空间中，支持向量数目大大减少，进而大大降低了测试集的错误率。

5 结 论

实验结果表明，采用主成分分析降维算法降维到 20 维的特征数据在线性 SVM 和 LSSVM 分类器上的平均识别率都高于 97%，最高识别率都可达到 100%，且模型稳定。采用主成分分析结合基于网格搜索优化的 LSSVM 方法在利用较低维数数据进行分类时，比线性 SVM 方法的识别精度高，稳定性好。本研究证明了用主成分分析结合线性 SVM 或 LSSVM 识别定性分析算法鉴别椴树蜜、荆条蜜、油菜蜜、洋槐蜜、荔枝蜜 5 种蜂蜜品种是可行的，而且鉴定过程中工作效率高，减少了不必要的损失，降低了鉴定成本；此外还能够避免传统鉴定方法中主观判断对鉴定结果造成的影响。本研究采用的 SVM 分类方法比较成熟，今后还会探讨其他较为成熟的分类器方法，以验证蜂蜜中红外光谱数据的分类效果。

参 考 文 献

[1] Chen L Z. Study on quality evaluation for honey by near infrared spectroscopy[D]. Beijing: Institute of Quality Standards and Testing Technology for AGRO-Products of CAAS, 2010.
陈兰珍. 蜂蜜品质近红外光谱评价技术研究[D]. 北京: 中国农业科学院农业质量标准与检测技术研究所, 2010.

[2] Liu B J. The research of detection method about characteristic of honey producing area[D]. Baoding: Hebei University, 2010.
刘博静. 蜂蜜产地特征检测方法的研究[D]. 保定: 河北大学, 2010.

[3] Zhong Y P, Zhong Z S, Chen L Z, *et al*. Qualitative identification of floral origin and adulteration of honey by near-infrared spectroscopy [J]. Modern Food Science and Technology, 2010, 26(11): 1280-1282.
钟艳萍, 钟振声, 陈兰珍, 等. 近红外光谱技术定性鉴别蜂蜜品种及真伪的研究[J]. 现代食品科技, 2010, 26 (11): 1280-1282.

[4] Ruoff K, Luginbühl W, Künzli R, *et al*. Authentication of the botanical and geographical origin of honey by mid-infrared spectroscopy [J]. Journal of Agricultural and Food Chemistry, 2006, 54(18): 6873-6880.

[5] Bertelli D, Plessi M, Sabatini A G, *et al*. Classification of Italian honeys by mid-infrared diffuse reflectance spectroscopy (DRIFTS) [J]. Food Chemistry, 2007, 101(4): 1565-1570.

[6] Graça G, Moreira A S, Correia A J , *et al*. Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: a possible avenue for early diagnosis of prenatal disorders?[J]. Analytica Chimica Acta, 2013, 764: 24-31.

[7] Zhang W J, Chen L Z, Wu L M, *et al*. Study of mid-infrared spectroscopy analysis for rapid discrimination of botanical origin of honey[C]. Summit Forum of National Bee Industry, 2013.
张文娟, 陈兰珍, 吴黎明, 等. 中红外光谱法快速鉴别不同蜜源蜂蜜[C]. 全国蜂产业高峰论坛, 2013.

[8] Hu Y Q, Yin C L, Ma W K, *et al*. Identification of adulterated honey based on infrared spectroscopy and pattern recognition technology[J]. Chinese Journal of Applied Chemistry, 2011, 28(s1): 144-145.
胡乐乾, 尹春玲, 马渭奎, 等. 红外光谱法对蜂蜜掺伪的模式识别[J]. 应用化学, 2011, 28(s1): 144-145.

[9] Sun Y, Zhang H H, Wang Z. Application of infrared spectrum technology in Raohe honey characterization of traceability[J]. Chemical Analysis and Meterage, 2015, 24(3): 41-44.
孙燕, 张海华, 王铮. 中红外光谱技术应用于饶河蜂蜜产地溯源的表征[J]. 化学分析计量, 2015, 24(3): 41-44.

[10] Duan F H, Wang X H, Ye H H, *et al*. Carbon dioxide retrieval method based on statistics and optical path distribution[J]. Acta Optica Sinica, 2017, 37(5): 0501003.
段锋华, 王先华, 叶函函, 等. 基于统计与光程分布的二氧化碳反演方法[J]. 光学学报, 2017, 37(5): 0501003.

[11] Cheng L Y, Mi G Y, Li S, *et al*. Quality diagnosis of joints in laser brazing based on principal component analysis-support vector machine model[J]. Chinese Journal of Lasers, 2017, 44(3): 0302004.
程力勇, 米高阳, 黎硕, 等. 基于主成分分析-支持向量机模型的激光钎焊接头质量诊断[J]. 中国激光, 2017, 44(3): 0302004.

[12] Liao J S, Wang L G. Hyperspectral image classification method based on fusion with two kinds of spatial information[J]. Laser & Optoelectronics Progress, 2017, 54(8): 081002.
廖建尚, 王立国. 两类空间信息融合的高光谱图像分类方法[J]. 激光与光电子学进展, 2017, 54(8): 081002.

[13] Chen L Z, Sun Q, Ye Z H, *et al*. Determination of floral origin of honey by near infrared spectroscopy based on artificial neural network[J]. Food Science of Technology, 2009, 34(8): 287-289.
陈兰珍, 孙谦, 叶志华, 等. 基于神经网络的近红外光谱鉴别蜂蜜品种研究[J]. 食品科技, 2009, 34(8): 287-289.

[14] Zhang Y N, Chen L Z, Xue X F, *et al*. Discrimination of rice syrup adulterant of acacia honey based using near-infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2015, 35(9): 2536-2539.
张妍楠, 陈兰珍, 薛晓锋, 等. 基于近红外光谱检测技术鉴别洋槐蜜中掺入大米糖浆的可行性研究[J]. 光谱学与光谱分析, 2015, 35(9): 2536-2539.

[15] Chen B M, Fan X P, Zhou Z M, *et al*. The principle and prospect of support vector machine [J]. Manufacturing Automation, 2010, 32(12): 136-138.
陈冰梅, 樊晓平, 周志明, 等. 支持向量机原理及展望[J]. 制造业自动化, 2010, 32(12): 136-138.

[16] Chen W H. Research on classification of hyperspectral images based on support vector machine [D]. Harbin: Harbin Engineering University, 2008.
陈万海. 基于支持向量机的超谱图像分类技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2008.

[17] Chen H Z, Chen F, Xu L L, *et al*. Grid search parameter optimization applied to near infrared LSSVM modeling quantitative analysis of fishmeal ash[J]. Journal of Analytical Science, 2016, 32(2): 198-202.
陈华舟, 陈福, 许丽莉, 等. 基于网格搜索的参数优化方法用于鱼粉灰分的近红外 LSSVM 定量分析[J]. 分析科学学报, 2016, 32(2): 198-202.

[18] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1998: 1-17.

[19] Duan K B, Rajapakse J C, Nguyen M N. One-Versus-One and One-Versus-Allmulticlass SVM-REF for gene selection in cancer classification [C] // Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Valencia: [s. n.], 2007: 47-56.

[20] Chang C C, Lin C J. LIBSVM: a library for support vector machines[EB/OL]. (2010-03-01) [2017-11-20]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[21] Tang X B. Seismic reservoir discrimination based on support vector machines [D]. Chengdu: Chengdu University of Technology, 2009.
唐小彪. 基于支持向量机的地震储层预测方法研究[D]. 成都: 成都理工大学, 2009.