

基于 KPCA 和 LSSVM 的蜂蜜近红外光谱鉴别分析

谈爱玲, 毕卫红

(燕山大学信息科学与工程学院光电子工程系, 河北 秦皇岛 066004)

摘要:为快速鉴别普通植物源与中草药植物源蜂蜜,提出一种核主成分分析和最小二乘支持向量机相结合的蜂蜜近红外光谱定性分析新方法。利用傅里叶变换近红外光谱仪测定普通洋槐蜂蜜和益母草、黄连两种中药植物源蜂蜜样本的近红外光谱并预处理,然后对光谱进行核主成分分析,提取非线性特征,最后设计基于纠错编码最小二乘支持向量机的多类分类器模型。采用网格搜索法确定模型最优参数,利用最优分类模型对未知类别蜂蜜样本进行识别,正确率可达 96.67%。结果表明,基于 KPCA 和 LSSVM 的近红外光谱定性分析算法鉴别普通植物源与中草药植物源蜂蜜是可行的。

关键词:近红外光谱;核主成分分析;最小二乘支持向量机;蜂蜜;中草药植物源

中图分类号:0657.3 **文献标识码:**A **DOI:**10.3969/j.issn.1001-5078.2011.12.009

Identification of honey by NIR spectroscopy technology based on KPCA and LSSVM

TAN Ai-ling, BI Wei-hong

(Institute of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: For the rapid identification of common plant honey and Chinese medicine nectar plant honey, a novel qualitative identification method combined Kernel Principal Component Analysis (KPCA) and Least Square Support Vector Machine (LSSVM) is proposed. The presented method uses Fourier transform NIR spectrophotometer to collect the spectral data of locust, leonuri and coptis honey. The KPCA algorithm is used to extract nonlinear features. Then a LSSVM classification model based on Error Correcting Output Code (ECOC) is designed and grid search method is used to determine the optimal model parameters. To unknown honey samples, the optimal model has the best identification capability with a accuracy of 96.67%. Experimental results indicate that the proposed qualitative analysis method based on KPCA and LSSVM can distinguish the honey of common and Chinese medicine nectar plant.

Key words: NIR spectroscopy; kernel principal component analysis; least square support vector machine; honey; Chinese medicine nectar plant

1 引言

蜂蜜是蜜蜂采集蜜源植物的花内或花外蜜腺,经过充分酿造加工而成。现代研究表明,蜂蜜是一种营养丰富的食疗佳品,其中含有单糖及少量的矿物质、维生素、蛋白质、有机酸、酶类等多种营养成分^[1]。我国是蜂蜜大国,各种蜜源植物有 500 多种,可以提供大量商品蜜粉源植物也有 40 余种,如北方

常见的洋槐蜜、椴树蜜、枣花蜜、南方的龙眼蜜、荔枝蜜、琵琶蜜等。这其中还有一些特种蜂蜜,所谓特种

基金项目:教育部高等学校科技创新工程重大项目培育基金(No. 708025);河北省自然科学基金(No. F2010001268);秦皇岛市科学技术研究与发展计划(No. 200901A032)资助。

作者简介:谈爱玲(1978-),女,讲师, E-mail: tanailing@ysu.edu.cn

收稿日期:2011-05-16; **修订日期:**2011-06-03

蜂蜜,即指蜜蜂从特别稀有的蜜源植物上采集酿造的天然蜂蜜。这类蜂蜜常为单一中草药植物花蜜,产量低,品质优,具有独特的芳香和特别的药用价值,比如益母草蜜、黄连蜜、丹参蜜等。目前市场上特种蜂蜜价格比普通蜂蜜高出一倍以上,仍然供不应求。一些不法商贩为谋取利益,以次充好,侵害消费者利益。因此,不同植物源蜂蜜,尤其是特种蜂蜜的准确、快速、无损检测具有非常重要的意义。

近红外光谱分析技术以其速度快、效率高、结果稳定、重复性好等优点已经在石油化工、农业、食品、医药等领域中得到广泛的应用^[2-6],尤其在蜂蜜的近红外光谱分析领域也有很多成功应用^[7-10]。文献[7]结合判别偏最小二乘法建立蜂蜜真伪鉴别的数学模型;文献[8]提出一种基于独立组分分析的可见/近红外光谱透射技术鉴别不同蜂蜜品牌的方法,鉴别准确率可达100%;文献[9]采用主成分分析结合贝叶斯线性判别和前向神经网络多分类器实现蜜源的快速无损识别;文献[10]采集蜂蜜近红外光谱,建立定量模型预测蜂蜜中果糖和葡萄糖含量。

主成分分析(principal component analysis, PCA)是特征提取与数据降维的常用方法^[11],核主成分分析(kernel principal component analysis, KPCA)是一种非线性主成分分析方法,充分利用核函数来解决非线性映射问题,具有很好的非线性逼近能力^[12]。文献[13]、[14]利用KPCA进行光谱信号的特征提取,取得了很好的效果。支持向量机(support vector machine, SVM)建立在统计学习理论的VC维理论和结构风险最小原则基础上,在样本的非线性、稀疏性和高维模式识别方面具有独特的优势^[15];Suykens在SVM的基础上提出了最小二乘支持向量机(least square support vector machines, LSSVM)^[16],LSSVM采用二次损失函数,将SVM中的二次规划问题转化为线性方程组求解,在保证精度的前提下大幅降低计算复杂性,加快求解速度。SVM和LSSVM算法在光谱分析领域也都有一些成功的应用^[17-19]。

本文提出一种结合核主成分分析和多类最小二乘支持向量机算法的中草药植物源与普通植物源蜂蜜近红外光谱特征提取和分类模型。首先测得一种普通蜂蜜和两种特殊单一中草药植物源蜂蜜的近红外光谱,并进行预处理;然后,利用核主成分

分析算法提取光谱的非线性特征;基于核主成分特征对多分类最小二乘支持向量机模型进行训练,通过基于交叉验证的网格搜索算法获得最优模型参数;最后对未知样本进行测试,验证提出算法的有效性。

2 材料与方法

2.1 实验仪器

实验使用德国布鲁克光学仪器公司(Bruker Optics Inc.)的MPA型傅里叶近红外光谱仪,该仪器的光谱扫描范围12000~4000 cm^{-1} ,扫描次数为64次,光谱分辨率为8 cm^{-1} ,TE-InGaAs检测器。分析软件为仪器配套OPUS 6.5。

2.2 样品制备与光谱测量

实验中蜂蜜样本选购自本地某超市,为汪氏品牌的洋槐蜜、益母草蜜和土黄连蜜三种蜂蜜样本。样本首先置于50 $^{\circ}\text{C}$ 的水浴中,搅匀,然后冷至25 $^{\circ}\text{C}$,恒温。每个品种蜂蜜采集50个样本,每个样本测量5次光谱后取平均值,最后得到3类共150个样本的光谱。三种蜂蜜样品的近红外光谱如图1所示。

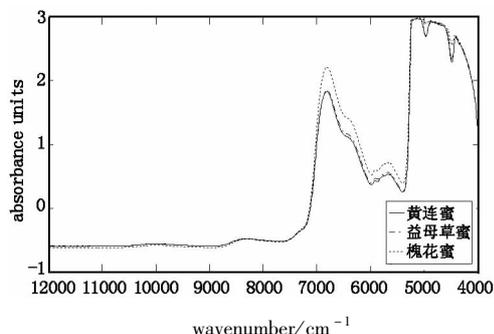


图1 三种蜂蜜样本的近红外光谱图

Fig. 1 near infrared spectra of three kinds of honey samples

从光谱图中可以看出在近红外全谱段每种蜂蜜都有四簇谱峰,由于蜂蜜是一种复杂的糖类混合物,成分复杂,每簇谱峰都可能是若干个不同基频的倍频和合频吸收的组合,谱带归属较为困难,但三种蜂蜜样品在7000~4500 cm^{-1} 区域有较为明显不同,本文基于此段光谱数据进行处理。在进行特征提取和分类之前,为去除噪声、散射等的影响,采用Savitsky-Golay10点平滑结合多元散射校正方法对原始近红外光谱进行预处理。

2.3 KPCA 特征提取

主成分分析是统计学中的一种有效的特征提取方法,但其只能提取线性特征来反映变量间的线性

关系。由于蜂蜜成分的复杂性,变量之间往往呈现出较强的非线性关系,因此就需要非线性方法对蜂蜜光谱信号进行分析。所以本文应用其非线性扩展方法 KPCA。其基本思想是通过非线性变换 $\phi(\cdot)$ 将样本数据从输入空间映射到高维特征空间 $F:\phi(x)$,然后在高维特征空间中对 $\phi(x)$ 进行线性 PCA 计算,就相当于在输入空间 x 中的非线性 PCA。

给定 N 个样本 $x_1, x_2, \dots, x_N \in R^m$, 由非线性函数 $\phi(\cdot)$ 将输入数据从原空间映射到高维特征空间 $F, \phi(x_j)$ 的协方差矩阵 C 为:

$$C^F = \frac{1}{N} \sum_{j=1}^N \phi(x_j) \phi(x_j)^T \quad (1)$$

式中,特征值和特征向量为: $\lambda V = C^F V$, 特征值 $\lambda \geq 0$; V 为特征向量。通过计算训练数据在特征向量 V 上的投影计算主成分。核主成分分析详细过程参考文献[9]。

2.4 ECOC-LSSVM 分类器

SVM 基本思想是通过定义适当的核函数实现非线性变换,将输入空间变换到一个高维空间,然后在这个新空间中求取最优线性分类面。最小二乘支持向量机是在支持向量机基础上将不等式约束转化成等式约束。

设有 n 个数据样本 $(x_i, y_i), (i = 1, \dots, n)$, 其中 x_i 为输入, y_i 为输出。LSSVM 可描述为如下优化问题:

$$\min_{w, b} J(w, \xi) = \frac{1}{2} \omega^T \omega + \frac{1}{2} f \sum_{i=1}^n \varepsilon_i^2$$

$$s. t. y_i [\omega^T \varphi(x_i) + b] = 1 - \varepsilon_i; i = 1, \dots, n \quad (2)$$

式中, J 为目标函数; ω 为权向量; b 为偏置; ε 为松弛变量,用来度量数据点对模式可分理想条件下的偏离程度, f 为平衡分类误差和算法复杂度的惩罚因子。非线性映射 $\varphi(x)$ 将样本 x 从原空间映射到更高维的特征空间。该优化问题对应的 Lagrange 方程为:

$$L(w, b, \xi, a) = J(w, \xi) - \sum_{i=1}^n a_i \{ [w^T \varphi(x_i) + b] - 1 + \xi_i \} \quad (3)$$

式中, a_i 为 Lagrange 乘子。根据优化条件: $\partial L / \partial w = 0, \partial L / \partial \xi_i = 0, \partial L / \partial b = 0, \partial L / \partial a_i = 0$, 消去 w 和 ξ 可得以下线性方程组:

$$\begin{bmatrix} 0 & I^T \\ I & K + f^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (4)$$

式中, $I = [1, 1, \dots, 1]_{1 \times n}^T, I$ 为单位矩阵。可解得 b

和 a , 则分类决策函数为:

$$f(x) = \text{sgn} \left[\sum_{i=1}^n a_i K(x, x_i) + b \right] \quad (5)$$

针对多分类问题,需要组合多个二类 SVM 分类器来构造 SVM 多分类器。常用的组合方法有:一对一 SVM,一对多 SVM,决策导向无环图 SVM,纠错输出编码法等^[20]。其中,纠错编码方法针对 M 类数据分类问题,对每个类进行长度为 L 的二进制编码,就把 M 类分类问题转化为 L 个两类分类问题。每个码位上采用 LSSVM 作为码位分类器。对于一个新样本, L 个 LSSVM 的分类结果构成一个码字 V , V 与哪个类别已知编码的汉明距离最小,则就属于哪个类别。纠错编码方法具有运算速度快,推广性强等优点^[21],本文采用该算法对三种蜂蜜进行识别。

根据上述 KPCA 特征提取和 ECOC-LSSVM 分类器原理,提出基于 KPCA-LSSVM 的识别方法,此方法充分结合两者的优点,具体的实现步骤如下:

(1) 将采集到的光谱分成训练集 V 和测试集 T , 针对 V 应用 KPCA 算法进行特征提取,得到蜂蜜光谱的低维非线性特征矢量,特征维数为 R ,组成用于光谱识别模型训练的特征向量 H ;

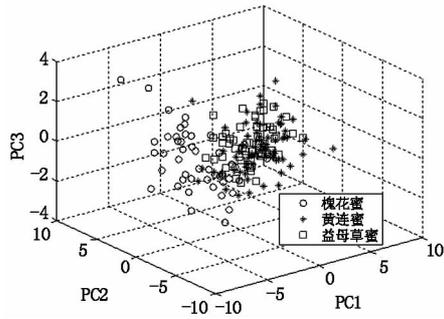
(2) 构造近红外光谱识别的多类 LSSVM 分类模型,利用 H 对模型训练,采用网格搜索法得到模型最优参数;

(3) 针对 T 经 KPCA 提取 R 维特征向量,作为已训练好的 LSSVM 模型的输入,其输出即为对应的未知蜂蜜样本定性分析结果。

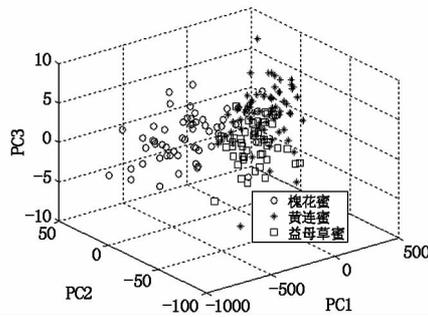
3 实验结果与分析

3.1 核主成分特征分析

分别应用 PCA 和 KPCA 特征提取方法对样本数据进行处理,其中前 3 个主成分的得分投影散点图分别如图 2(a) 和图 2(b) 所示。对比可以看出,普通蜂蜜和特种蜂蜜有较好的可分性,这点从光谱上也可以反映出来;利用 PCA 线性特征提取方法很难区分黄连和益母草两种特种蜂蜜,而利用 KPCA 非线性特征提取方法能够较好的区分两种特种蜂蜜;同时给出前 5 个主成分累积贡献率,如表 1 所示, KPCA 方法前 5 个主成分的累计贡献率已经达到了 99.98%, 同时 KPCA 方法用 3 个主成分表达的信息就超过了 PCA 方法 5 个主成分的效果,这说明 KPCA 方法用更少的主成分就能更全面的反映全部光谱包含的信息,优于传统 PCA 方法。



(a) 主成分方法
(a) PCA



(b) 核主成分方法
(b) KPCA

图 2 蜂蜜样本前三个主成分得分投影图
Fig.2 score plot of the top three principal components for honey samples

表 1 PCA 和 KPCA 主成分贡献率对比

Tab.1 comparison of PC accumulative ratio of contribution between PCA and KPCA method

	PC1	PC2	PC3	PC4	PC5
PCA/%	86.75	93.12	95.13	96.78	98.23
KPCA/%	92.12	96.98	98.54	99.46	99.98

3.2 支持向量机参数优化

设计 LSSVM 分类器时,关键是其结构和参数的选择。其中核函数选择一般依赖于特定数据,通过实验的方法选取和确定,没有固定的原则。本文分别选取多项式函数、径向基函数 RBF 和 Sigmoid 函数作为核函数进行验证,结果发现 RBF 作为核函数的分类精度较高,最终确定核函数为 RBF 函数。

RBF 核函数主要涉及两个参数: γ 和 σ^2 。 γ 为惩罚参数,它控制对错分样本的惩罚程度,用来保持样本偏差与机器泛化能力之间的平衡。径向基核宽度 σ^2 是另一个重要参数, σ^2 值太小或太大,会对样本数据造成过学习或欠学习的现象;本文首先确定惩罚参数和核参数对 (γ, σ^2) ,然后采用网格搜索法从参数集中选取不同参数组合,对最小二乘支持向量机进行训练,从参数组合中选择识别率最高对应的一个参数组合作为最优的 γ 和 σ^2 ,参数值为(16,

9),具体结果如图 3 所示。

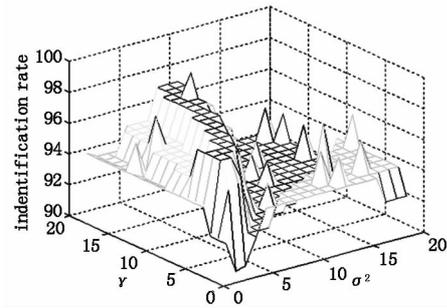


图 3 调整 γ 和 σ^2 确定最优 LSSVM 模型
Fig.3 tuning parameter γ and σ^2 to identify the optimal LSSVM model

3.3 定性分类结果

算法采用 MATLAB R2008a 编程,应用 LSSVM 工具箱^[22],在 P4 2.66 GHz,512 MB 内存 PC 机上运行。针对 3 类蜂蜜共 150 个样本,其中 105 个样本用来训练分类器,剩余的 45 个样本做测试集。将提出的 KPCA + LSSVM 方法与经典的 PCA + SVM 方法,以及 KPCA + SVM 方法对比考察提出方法的性能,识别结果如表 2 所示。由于蜂蜜成分复杂,应用 PCA 算法只能提取线性特征,不能很好地反映变量间的非线性关系。而 KPCA 方法能够有效的处理非线性关系,因此采用 KPCA 特征提取比采用 PCA 提取特征变量后进行分类的总体效果好;同样基于 KPCA 特征,LSSVM 分类和 SVM 分类算法都取得较好的分类效果,但是 LSSVM 算法的执行效率更高。

表 2 三种方法识别结果

Tab.2 identification results of three methods

方法	训练集/%	测试集/%
PCA + SVM	91.11	86.67
KPCA + SVM	95.56	95.56
KPCA + LSSVM	97.78	95.56

4 结论

蜂蜜样本成分非常复杂,KPCA 作为传统 PCA 特征提取方法的扩展,可以更好地揭示其变量之间的非线性关系;LSSVM 相比传统 SVM 算法复杂度大为降低,满足商品质量监督部门对于蜂蜜种类鉴别的快速性要求;纠错编码多分类原理具有一定的错误容忍度,可解决蜂蜜样本检测现场采集过程中的强噪声问题。本文结合三者的优势,建立了基于 KPCA 特征提取结合 ECOC-LSSVM 分类器的不同植物源蜂蜜近红外光谱定性分析模型,通过寻优算法确定了模型的最优参数,针对三种典型蜂蜜样本的

实验表明该方法能够取得理想的效果。

参考文献:

- [1] GB 18796 - 2005. Honey [S]. National Standards of the People's Republic of China. (in Chinese)
GB 18796 - 2005. 蜂蜜 [S]. 中国.
- [2] Yan Yanlu, Zhao Longlian, Han Donghai, et al. Foundation and application of near-infrared spectroscopy analysis [M]. Beijing: China Light Industry Press, 2005. (in Chinese)
严衍禄, 赵龙莲, 韩东海, 等. 近红外光谱分析基础与应用 [M]. 北京: 中国轻工业出版社, 2005.
- [3] Lu Wanzhen, Yuan Hongfu, Xu Guangtong, et al. Modern near infrared spectroscopy analytical technology [M]. 2nd ed. Beijing: Chinese Petrochemical Industry Press, 2007. (in Chinese)
陆婉珍, 袁洪福, 徐广通, 等. 现代近红外光谱分析技术 [M]. 2 版. 北京: 中国石化出版社, 2007.
- [4] Liang Liang, Yang Minhua, Liu Zhixiao. Discrimination of lines and authenticity of hybrid rice seed with visible-near infrared spectra [J]. Laser & Infrared, 2009, 39 (4): 407 - 410. (in Chinese)
梁亮, 杨敏华, 刘志霄, 等. 杂交稻种品的系与真伪的可见 - 近红外光谱鉴别 [J]. 激光与红外, 2009, 39 (4): 407 - 410.
- [5] Zhang Xiaohui, Liu Jianxue. Identification of forsythia suspense from different habitats by NIR spectra [J]. Laser & Infrared, 2008, 38 (4): 342 - 344. (in Chinese)
张晓慧, 刘建学. 近红外光谱技术鉴别连翘产地 [J]. 激光与红外, 2008, 38 (4): 342 - 344.
- [6] Yan Wenjuan, Lin Ling, Zhao Jing, et al. Probability neural network for the classification of tongue diagnosis by near infrared spectroscopy [J]. Laser & Infrared, 2010, 40 (11): 1201 - 1204. (in Chinese)
严文娟, 林凌, 赵静, 等. 概率神经网络用于舌诊的近红外光谱分类 [J]. 激光与红外, 2010, 40 (11): 1201 - 1204.
- [7] Chen Lanzhen, Zhao Jing, Ye Zhifeng, et al. Determination of adulteration in honey using near-infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2008, 28 (11): 2565 - 2567. (in Chinese)
陈兰珍, 赵静, 叶志华, 等. 蜂蜜真伪的近红外光谱鉴别研究 [J]. 光谱学与光谱分析, 2008, 28 (11): 2565 - 2567.
- [8] Shao Yongni, He Yong, Bao Yidan, et al. Application of visible/near infrared spectroscopy to discriminating honey brands based on independent component analysis and BP neural network [J]. Spectroscopy and Spectral Analysis, 2008, 28 (3): 602 - 604. (in Chinese)
邵咏妮, 何勇, 鲍一丹. 基于独立组分析法和 BP 神经网络的可见/近红外光谱蜂蜜品牌的鉴别 [J]. 光谱学与光谱分析, 2008, 28 (3): 602 - 604.
- [9] Yang Yan, Nie Pengcheng, Yang Haiqing, et al. Rapid recognition method of nectar plant based on visible-near infrared spectroscopy [J]. Transactions of the Chinese Society of Agricultural Engineering, 2010, 26 (3): 238 - 242. (in Chinese)
杨燕, 聂鹏程, 杨海清, 等. 基于可见 - 近红外光谱技术的蜜源快速识别方法 [J]. 农业工程学报, 2010, 26 (3): 238 - 242.
- [10] Tu Zhenhua, Zhu Dazhou, Ji Baoping, et al. Difference analysis and optimization study for determination of fructose and glucose by near infrared spectroscopy [J]. Chinese Journal of Analytical Chemistry, 2010, 38 (1): 45 - 50. (in Chinese)
屠振华, 朱大洲, 籍保平, 等. 蜂蜜中果糖和葡萄糖近红外检测的差异性分析及优化研究 [J]. 分析化学, 2010, 38 (1): 45 - 50.
- [11] Johnson R A, Wichern D W. Applied multivariate statistical analysis [M]. 6th ed. Beijing: Tsinghua University Press, Johnson R A, Wichern D W, 2008. (in Chinese)
实用多元统计分析 [M]. 6 版. 北京: 清华大学出版社, 2008.
- [12] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10 (5): 1299.
- [13] Weng Xinxin, Zhang Zhonghu, Yin Lihui, et al. Rapid determination of hypoglycemic tablets by handheld raman spectrometer and KPCA-clustering analysis [J]. Spectroscopy and Spectral Analysis, 2010, 30 (4): 984 - 987. (in Chinese)
翁欣欣, 张中湖, 尹利辉, 等. KPCA - 聚类分析法和用便携式拉曼仪快速鉴别降糖药 [J]. 光谱学与光谱分析, 2010, 30 (4): 984 - 987.
- [14] Hao Huimin, Tang Xiaojun, Bai Peng, et al. Spectroscopy and spectral analysis [J]. Quantitative Analysis of Multi-Component Gas Mixture Based on KPCA and SVR, 2008, 28 (6): 1286 - 1289. (in Chinese)
郝惠敏, 汤晓君, 白鹏, 等. 基于核主成分分析和支持向量回归机的红外光谱多组分混合气体定量分析 [J]. 光谱学与光谱分析, 2008, 28 (6): 1286 - 1289.
- [15] Vapnik V N. The Nature of Statistical Learning [M]. New York: Springer, 1995.
- [16] Suykens J A K, Vandewalle J. Neural Network Letters (S105727122) [J]. 1999, 19 (3): 293.
- [17] Ma Chaojie, Li Xiaoxia, Yang Hua, et al. Multi-view target

- recognition algorithm based on support vector machine classification[J]. *Laser & Infrared*, 2009, 39 (1): 88 - 91. (in Chinese)
- 马超杰, 李晓霞, 杨华, 等. 应用支持向量机分类的多角度目标识别技术[J]. *激光与红外*, 2009, 39 (1): 88 - 91.
- [18] Liu Fei, Wang Li, He Yong, et al. Detection of spad value of cucumber leaves based on visible/near infrared spectroscopy technology[J]. *Journal of Infrared and Millimeter Waves*, 2009, 28 (4): 272 - 276. (in Chinese)
- 刘飞, 王莉, 何勇, 等. 基于可见/近红外光谱技术的黄瓜叶片 SPAD 值检测[J]. *红外与毫米波学报*, 2009, 28 (4): 272 - 276.
- [19] Lin Hao, Zhao Jiewen, Chen Quansheng, et al. Identification of egg freshness using near infrared spectroscopy and one class support vector machine algorithm[J]. *Spectroscopy and Spectral Analysis*, 2010, 30 (4): 929 - 932. (in Chinese)
- 林颢, 赵杰文, 陈全胜, 等. 近红外光谱结合一类支持向量机算法检测鸡蛋的新鲜度[J]. *光谱学与光谱分析*, 2010, 30 (4): 929 - 932.
- [20] Chih W H, Chih J L. A Comparison of methods for multi-class support vector machines [J]. *Neural Networks*, 2002, 13(2): 415.
- [21] Dietterich T G, Bakiri G. Solving multiclass learning problems via error-correcting output codes[J]. *Journal of Artificial Intelligence Research*, 1995, (2): 263.
- [22] <http://www.esat.kuleuven.be/sista/lssvmlab/>.