基于 PCA-Stacking 模型的食源性致病菌拉曼光谱识别

史如晋[1]，夏钒曾[2]，曾万聃[1*]，曲晗[3]

[1]上海应用技术大学计算机科学与信息工程学院，上海 201418；

[2]吉林大学软件学院，吉林 长春 130122；

[3]军事医学科学院军事兽医研究所吉林省人畜共患病预防与控制重点实验室，吉林 长春 130122

摘要 食源性致病菌的快速识别是一项重要的工作，与传统检测方法相比，拉曼光谱能在无损检测的同时加快鉴别速度。为了提高大肠杆菌 O_{157}∶H_7 以及布鲁氏菌 S2 株拉曼光谱识别的准确性和效率，提出一种基于主成分分析与 Stacking 算法的集成判别模型，使用网格搜索以及 K 折交叉验证来提高模型的稳健性。与逻辑回归、K 近邻、支持向量机等单一模型进行对比，实验结果证明 PCA-Stacking 集成模型有最高的准确率，达 99.73%，达到了预期效果。

关键词 光谱学；拉曼光谱；机器学习；Stacking 模型；食源性致病菌

中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.043003

Raman Spectroscopic Classification of Foodborne Pathogenic Bacteria Based on PCA-Stacking Model

Shi Rujin[1], Xia Fanzeng[2], Zeng Wandan[1*], Qu Han[3]

[1] *School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai* 201418, *China*;

[2] *College of Software, Jilin University, Changchun, Jilin* 130122, *China*;

[3] *Jilin Provincial Key Laboratory for Disease Prevention and Control, Institution of Military Veterinary, Academy of Military Medical Sciences, Changchun, Jilin* 130122, *China*

Abstract The rapid identification of foodborne pathogenic bacteria is an important task. Compared with the traditional detection methods, Raman spectroscopy is a non-destructive testing method and can simultaneously enhance the identification speed. In order to improve the accuracy and efficiency of Raman spectroscopic identification of *Escherichia coil* O_{157}∶H_7 and *Brucella suis* vaccine strain S2, a integral classification model is proposed based on the principal component analysis and the Stacking algorithm, whose robustness is improved by the grid search and K-fold cross validation. It is experimentally confirmed that compared with the logistic regression, K nearest neighbor, support vector machine and other single models, the integral model based on the Stacking algorithm possesses the highest accuracy rate of 99.73% the expected result is achieved.

Key words spectroscopy; Raman spectroscopy; machine learning; Stacking model; foodborne pathogenic bacteria

OCIS codes 300.6450; 330.6230; 240.6695

1 引 言

食品安全问题是全世界共同关注的公共卫生话题，由食源性细菌引起的疾病已成为危害食品安全的最主要原因之一。目前，用于检测致病菌的方法有：形态学鉴定、免疫学检测及聚合酶链式反应[1-4](PCR)等。但是，这些方法操作步骤复杂、周期长，不能有效地起到监测和预防作用。

拉曼光谱是基于光和材料内化学键的相互作用而产生的，通过对拉曼光谱信号的分析，可对样品实

收稿日期：2018-06-27；**修回日期**：2018-08-11；**录用日期**：2018-09-06

基金项目：国家重点研发计划(2016YFC1201605)

* **E-mail**：zengwd@sit.edu.cn

现定性分析与定量计算。高玮村等[5]利用表面增强拉曼技术，通过人工识峰，成功检测出5种食源性致病菌。王宇田等[6]通过人工识峰，实现了对大肠杆菌 O_{157}∶H_7 的快速检测。何欣龙等[7]通过K近邻算法实现了塑钢窗的识别与分类。郭利斌[8]使用改进的支持向量机结合拉曼光谱实现了对癌症组织的分类与判别。

在没有先验知识的情况下，人工识谱分析会出现较大的误差，并且缺乏科学的识别评价标准。单一的机器学习分类算法与拉曼光谱分析结合在一起，虽然能降低操作者的工作量，在一定程度上提高检测的效率和可靠性，但是相比较而言，单一分类器的泛化性能弱于集成算法。

针对两种拉曼峰相似的食源性致病菌——大肠杆菌 O_{157}∶H_7 以及布鲁氏菌S2株，提出一种基于PCA-Stacking的集成分类算法。尝试寻求稳健性更好的数学统计模型和计算方法，针对拉曼光谱中存在的毛刺、基线漂移等问题，采用Savitzky-Golay滤波器以及非对称最小二乘实现光谱的预处理。结合网格搜索以及K折交叉验证，同时对相关结果进行讨论，证明了PCA-Stacking相比K近邻、支持向量机等单一分类器有更高的分类精度。

2 数据采集与预处理

2.1 实验样本

大肠杆菌 O_{157}∶H_7（CICC：21530）购于中国工业微生物菌种保藏管理中心（CICC），布鲁氏菌S2株由吉林省人畜共患病预防和控制重点实验室保存。

2.2 实验器材及过程

检测前将拉曼光谱仪（LabRAM HR Evolution，HORIBA Scientific）使用硅片（Si）在520.7 cm^{-1}的峰作为基准峰进行仪器校正。激发光源为632.8 mm的氦氖激光，激光强度为14 mW，20倍目镜，积分时间为3 s，积分2次，狭缝宽度为100 μm，测量范围为600～2000 cm^{-1}，分辨率为1 cm^{-1}。分别取5 μL待测样品滴于凹载玻片中央，使用LabSpec6.0软件进行光谱采集，每个样品采集60次，经筛选后得到52个大肠杆菌和54个布鲁氏菌拉曼光谱，共106个样本。

2.3 光谱预处理

大肠杆菌 O_{157}∶H_7 和布鲁氏菌S2株的原始拉曼光谱如图1所示。二者的光谱曲线在750，1129，1331，1464，1582 cm^{-1}左右都有相似的拉曼峰。从图1中可以看出，原始光谱中掺杂了荧光背景、毛刺等无用信息，这些噪声的存在对后续分析将会造成很大影响，故需对原始光谱进行相关预处理[9]。预处理过程包含对光谱进行归一化[10]、Savitzky-Golay平滑[11-12]以及非对称最小二乘扣除荧光背景，如图2所示，其中图2(a)为采集到的原始光谱数据，光谱图上存在很多噪声；图2(b)为经过Savitzky-Golay平滑后的光谱，毛刺现象几乎消除；图2(c)为去除荧光后的光谱图。

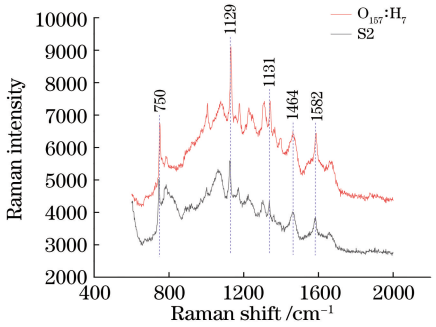

图1 原始拉曼光谱

Fig. 1 Original Raman spectra

3 光谱特征抽取

核酸、蛋白质、脂类和糖类均可生成独特的拉曼光谱，这是致病菌分析鉴定的重要依据。本研究中拉曼光谱的测量范围为600～2000 cm^{-1}，在维度如此高的情况下会出现数据样本稀疏、距离计算困难等问题。为缓解维数灾问题，使用主成分分析[13]实现对光谱特征的抽取。

对预处理后的106组光谱数据进行主成分分析，得到它们的帕累托图，如图3所示，其中横坐标代表主成分个数，纵坐标代表主成分的贡献率。从图3中可以发现，保留2个主成分和3个主成分后它们的贡献率分别达到91.24%和95.41%，即保留3个主成分几乎包含了所有的拉曼光谱信息。

抽取特征后的拉曼光谱在三维（3D）空间中具有很好的区分性，其中红色样点代表大肠杆菌 O_{157}∶H_7，蓝色样点代表布鲁氏菌S2株，它们在三维空间的具体分布如图4所示。

4 实验结果与讨论

4.1 K近邻（KNN）算法

KNN算法比较直观[14]。假设给定一个训练数据集，其中的实例类别已定。在分类时，对新的实例，根据其K个最相近的训练实例的类别，通过多数表决等方式进行预测。

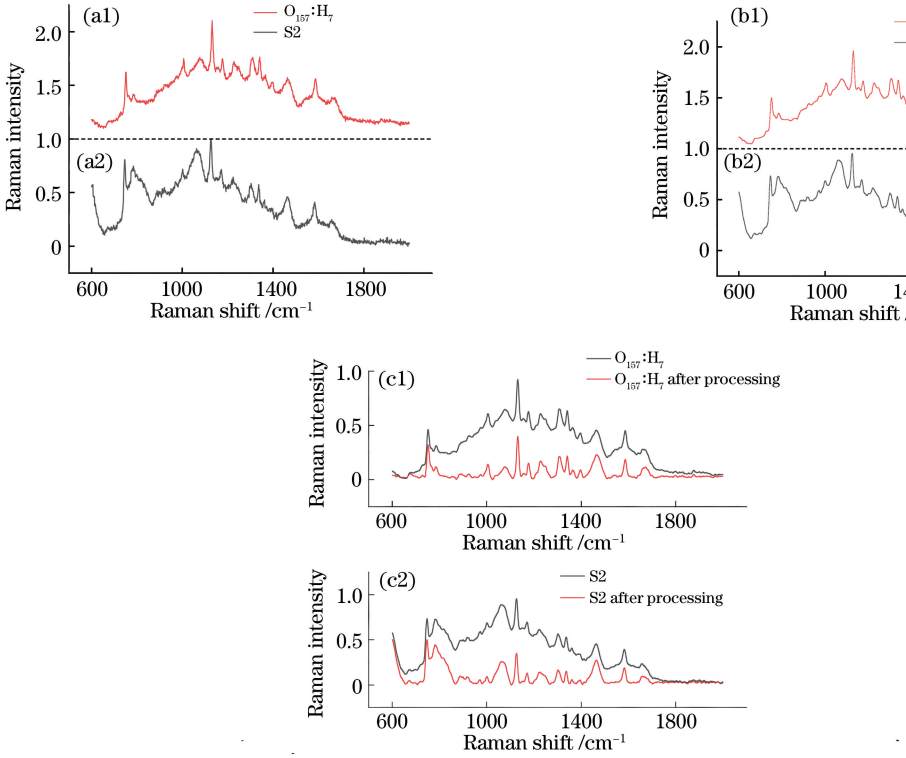

图 2 拉曼光谱预处理效果图。(a)归一化;(b)平滑去噪;(c)背景扣除

Fig. 2 Preprocessing effects of Raman spectra. (a) Normalization; (b) smoothing and denoising; (c) background deduction

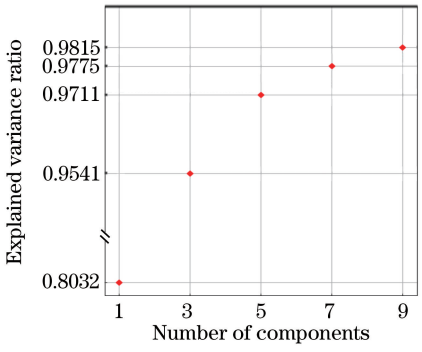

图 3 主成分帕累托图

Fig. 3 Preto chart of principle component

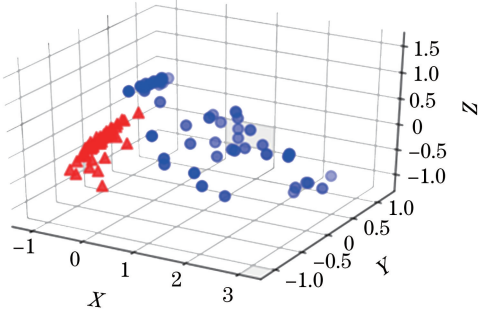

图 4 三维空间特征分布图

Fig. 4 Feature distribution in 3D space

K 值的选择、距离度量以及分类决策规则(往往是多数表决)是 K 近邻法的三个基本要素,它们将会影响分类的性能。将预处理好的数据集随机划分 30%作为测试集,70%作为训练集,K 值设定在 1 和 5 之间,候选度量距离为“曼哈顿距离”和“欧氏距离”。将上述对象作为网格搜索参数,并作十折交叉验证训练模型,在 K 为 2,以曼哈顿距离作为度量标准时,该模型最优分类准确率达 96.85%。

4.2 逻辑回归(LR)算法

二项逻辑回归模型[15]由条件概率分布 $P(Y|X)$表示,这里随机变量 X 取值为实数,随机变量 Y 取值为 1 或 0。这里,对逻辑回归模型所做的假设为

$$P(Y=1\mid x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)},\tag{1}$$

$$P(Y=0\mid x)=\frac{1}{1+\exp(w\cdot x)},\tag{2}$$

式中:$w\in\mathbf{R}$ 为参数;$w\cdot x$ 为 w 和 x 的内积。

对于模型参数的求解,使用极大似然估计(MLE),即找到一组参数,使得在这组参数下,似然度最大。设

$$P(Y=1\mid x)=\pi(x), \tag{3}$$

似然函数为

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}, \tag{4}$$

对数似然函数为

$$L(w)=\sum_{i=1}^{N}\{y_i\ln\pi(x_i)+(1-y_i)\ln[1-\pi(x_i)]\}=\sum_{i=1}^{N}\left\{y_i\ln\frac{\pi(x_i)}{1-\pi(x_i)}+\ln[1-\pi(x_i)]\right\}=\sum_{i=1}^{N}\{y_i(w\cdot x_i)-\ln[1+\exp(w\cdot x_i)]\}。\tag{5}$$

对 $L(w)$求极大值，即可得到 w 的估计值。

与 3.1 节中 KNN 使用的方法一样，将数据集送入 LR 模型中训练。相比于 KNN，逻辑回归在性能上有一定的改善，它的分类准确率达 97.21%。

4.3 支持向量机(SVM)算法

支持向量机的核心思想是在特征空间上寻找几何间隔最大的最优分离超平面，学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题。

假设在二维平面中有两类样本，如图 5 所示。其中，H_1 不能把类别分开；H_2 可以，但是只有很小的间隔；H_3 以最大的间隔把它们分开，即 H_3 对于样本局部扰动的“容忍”性最好，对未见示例的泛化性能最强。

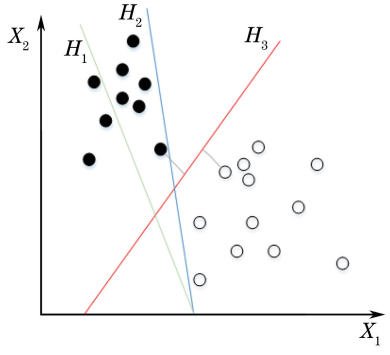

图 5 支持向量机示例

Fig. 5 Demonstration of SVM

与上述模型训练过程一样，将数据集送入 SVM 中训练，SVM 模型中网格搜索的对象为：错误项的惩罚参数 C 以及内核。在十折交叉验证后得到该模型的最佳参数：错误项惩罚参数 $C=0.1$，内核为线性核，该模型的分类准确率为 98.94%。

4.4 Stacking 算法

Stacking 算法也即 Stacked Generlization[16-17]，是一种集成学习模型，与单一模型相比，该方法可以提供更好的预测结果。Stacking 算法可以描述为：通过元分类器(Meta-Classifier)或元回归(Meta-Regresser)聚合多个分类或回归模型。基础层次模型(Level models)基于完整的训练集进行训练，然后元模型((Meta models)基于基础层次模型的输出进行训练，其架构如图 6 所示。

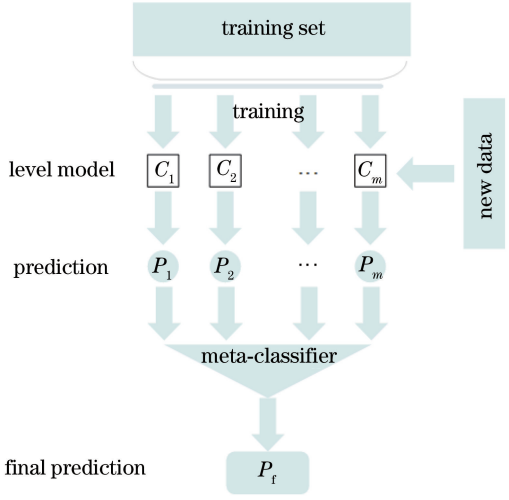

图 6 Stacking 算法架构图

Fig. 6 Flow chart of Stacking algorithm

图 6 中，$\{C_1,C_2,\cdots,C_m\}$ 为不同的学习分类器，因此 Stacking 通常是异构的。$\{P_1,P_2,\cdots,P_m\}$ 为分类器 $\{C_1,C_2,\cdots,C_m\}$ 的输出，P_f 为整个模型的输出。表 1 为算法的工作流程，用 $D=\{x_i,y_i\}_{i=1}^{m}$ 代表训练样本，H 代表集成分类器，h_t 代表集成分类器，$D_h=\{x'_i,y_i\}$ 为基础层次分类器产生的新数据集合，具体如下。

第一步：根据原始训练数据集学习得到基础层次的分类器。

第二步：根据基础层次分类器的输出构建新的数据集。在这一步中，基础层次分类器的输出被当作是新的特征。

第三步：根据新构建的特征学习得到元分类器。

Stacking 模型的基础层次模型由 KNN 和 SVM 组成，将 LR 作为元分类器。将预处理好的数据集随机划分 30%作为测试集，70%作为训练集，并作十折交叉验证，分类精确率达 99.73%。

与表现性能最差的 KNN 模型相比，Stacking 模型的分类准确度提高了 2.88%。这是因为 Stacking 算法是一种集成模型，从单一模型出发，反复学习，然后组合这些弱分类器的输出，构成一个强分类器，因此具有更高的可靠性。

表 1 Stacking 算法工作流程

Table 1 Flow chart of Stacking algorithm

Stacking algorithm
Input: Training data $D=\{x_i,y_i\}_{i=1}^{m}(x_i\in\mathbf{R}^n,y_i\in\{0,1\})$
Output: An ensemble classifier H
1: Step 1: Learn first-level classifiers
2: **for** $t\leftarrow1$ to T **do**
3: Learn a base classifier h_t based on D
4: **end for**
5: Step 2: Construct new data sets from D
6: **for** $i\leftarrow1$ to m **do**
7: Construct a new data set that contains $D_h=\{x'_i,y_i\}$ where $x'_i=\{h_1(x_i),\cdots,h_T(x_i)\}$
8: **end for**
9: Step 3: Learn a second-level classifier
10: Learn a new classifier h' based on the newly constructed data set
11: **return** $H(x)$

5 结 论

通过对比实验证明，对于大肠杆菌 O_{157}：H_7 以及布鲁氏菌 S2 株拉曼光谱，PCA-Stacking 集成模型具有最高的分类识别准确率。针对拉曼光谱的特殊性，对原始数据进行了平滑、去噪、荧光背景扣除以及降维等一系列预处理工作。此外，使用网格搜索以及交叉验证来确定模型的最佳参数。

由于样本数量不是很大，在建立分类模型时可能存在一定的过拟合。此外，实验中只研究了两种食源性致病菌的分类与识别，但方法具有普遍意义。后期将会在扩大实验样本的同时研究多种食源性致病菌的分类，构建相对完整的拉曼光谱数据库。

参 考 文 献

[1] Teng Y H, Suo B, Ai Z L, *et al*. Establishment and application of a multiplex PCR assay for simultaneous detection of *salmonella spp*. and *staphylococcus aureus* in quick-frozen foods[J]. Food Science, 2013, 34(8): 140-144.
滕要辉, 索标, 艾志录, 等. 速冻食品中沙门氏菌和金黄色葡萄球菌多重 PCR 检测方法的建立与应用[J]. 食品科学, 2013, 34(8): 140-144.

[2] Kim J S, Lee G G, Park J S, *et al*. A novel multiplex PCR assay for rapid and simultaneous detection of five pathogenic bacteria: *Escherichia coli* O_{157} : H_7, *Salmonella*, *Staphylococcus aureus*, *Listeria monocytogenes*, and *Vibrio parahaemolyticus*[J]. Journal of Food Protection, 2007, 70(7): 1656-1662.

[3] Zhang Y, Zhu L Q, Zhang Y T, *et al*. Simultaneous detection of three foodborne pathogenic bacteria in food samples by microchip capillary electrophoresis in combination with polymerase chain reaction[J]. Journal of Chromatography A, 2018, 1555: 100-105.

[4] Gong Q, Li Z L, Niu M F. A pilot study on PCR-based detection of four foodborne pathogenic microorganisms[J]. Journal of Food Measurement and Characterization, 2018, 12(2): 675-682.

[5] Gao W C, Li B, Wang X W, *et al*. Quick detection of five foodborne pathogenic bacteria based on surface enhanced Raman spectroscopy[J]. Journal of Jilin Agricultural University, 2017, 39(6): 733-737.
高玮村, 李博, 王习文, 等. 基于表面增强拉曼技术快速检测 5 种食源性致病菌[J]. 吉林农业大学学报, 2017, 39(6): 733-737

[6] Wang Y T, Qu H, Hao L Y, *et al*. Devising a rapid and efficient method of detecting *Escherichia coli* O_{157} : H_7 based on aptamer-mediated surface-enhanced Raman spectroscopy(SERS) [J]. Journal of Pathogen Biology, 2018, 13(1): 16-21.
王宇田, 曲晗, 郝良玉, 等. 基于核酸适配体 SERS 技术快速检测大肠埃希菌 O_{157}：H_7 的研究[J]. 中国病原生物学杂志, 2018, 13(1): 16-21.

[7] He X L, Chen L B, Wang J F, *et al*. Raman spectroscopy analysis of plastic steel window based on K nearest neighbors algorithm[J]. Laser & Optoelectronics Progress, 2018, 55(5): 053001.
何欣龙, 陈利波, 王继芬, 等. 基于 K 近邻算法的塑钢窗拉曼光谱分析[J]. 激光与光电子学进展, 2018, 55(5): 053001.

[8] Guo L B, Chen G N, Liu M Y, Analysis of biological tissue Raman spectroscopy data based on support vector machine algorithm [J]. Fu Lighting Technology, 2014, 25(2): 25-27.
郭利斌, 陈冠楠, 刘明宇. 基于支持向量机算法的生物组织拉曼光谱数据分析[J]. 福光技术, 2014, 25(2): 25-27.

[9] Zheng J W, Yang T W. Classification method of biological tissues based on Raman spectrum features [J]. Laser & Optoelectronics Progress, 2017, 54(5): 053001.
郑家文, 杨唐文. 基于拉曼光谱特征的生物组织识别方法[J]. 激光与光电子学进展, 2017, 54(5): 053001.

[10] Zhang H, Wang Q J, Zhu J J, *et al*. Influence of sample data preprocessing on BP neural network-based GPS elevation fitting[J]. Journal of Geodesy and Geodynamics, 2011, 31(2): 125-128.
张昊, 王琪洁, 朱建军, 等. 样本数据预处理对基于BP神经网络的GPS高程拟合的影响[J]. 大地测量与地球动力学, 2011, 31(2): 125-128.

[11] 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.

[12] Ma R, Wang Q, Chu D Z, *et al*. Study on a photoelectric signal processing method for the DOC online analyzer[J]. Journal of Ocean Technology, 2016, 35(6): 44-49.
马然, 王茜, 褚东志, 等. 一种DOC在线分析仪光电信号处理方法[J]. 海洋技术学报, 2016, 35(6): 44-49.

[13] Liu X H, Tan Q P, Zeng P, *et al*. Comparison and implementation of several MOOC-based text classification[J]. Computer Engineering & Software, 2016, 37(9): 27-33.
刘鑫昊, 谭庆平, 曾平, 等. 几种基于MOOC的文本分类算法的比较与实现[J]. 软件, 2016, 37(9): 27-33.

[14] Song Limei, Luo J. Pattern recognition[M]. Beijing: China Machine Press, 2015.
宋丽梅, 罗菁. 模式识别[M]. 北京: 机械工业出版社, 2015.

[15] Wei X P, Yu X C, Tan X, *et al*. A classification algorithm for hyperspectral images based on import vector machine[J]. Journal of Geomatics Science and Technology, 2015, 32(4): 379-383.
魏祥坡, 余旭初, 谭熊, 等. 一种基于输入向量机的高光谱影像分类算法[J]. 测绘科学技术学报, 2015, 32(4): 379-383.

[16] Wolpert D H. Stacked generalization [J]. Neural Networks, 1992, 5(2): 241-259.

[17] Aggarwal C C. Data classification: Algorithms and applications[M]. Boca Raton: CRC Press, 2014.