多波长透射光谱特征提取结合支持向量机的水体细菌识别方法研究

冯 春[1,2,3]，赵南京[1,3*]，殷高方[1,3*]，甘婷婷[1,3]，陈晓伟[1,2,3]，
陈 敏[1,2,3]，华 卉[1,2,3]，段静波[1,3]，刘建国[1,3]

1. 中国科学院环境光学与技术重点实验室，中国科学院安徽光学精密机械研究所，安徽 合肥 230031
2. 中国科学技术大学，安徽 合肥 230026
3. 安徽省环境光学监测技术重点实验室，安徽 合肥 230031

摘 要 实现水体致病菌的快速识别检测对防控由水体微生物污染引起的大规模疾病爆发有重要的现实意义。生化鉴定、核酸检测等常规细菌检测方法存在耗费时间长、需要精密的实验仪器等特点，不足以满足水体细菌微生物的快速实时在线监测。由于细菌的多波长透射光谱包含较丰富的特征信息，并且这项光谱检测技术具有快速简便、无接触、无污染等优点，近年来成为细菌检测研究的热点。以肺炎克雷伯氏菌、金黄色葡萄球菌、鼠伤寒沙门氏菌、铜绿假单胞菌和大肠埃希氏菌为研究对象，通过对细菌光谱作归一化处理和方差分析得到光谱变动最显著的特征波长区间，在该区间提取200 nm处的吸光度值及短波段的斜率值作为光谱特征值，结合支持向量机对不同种类细菌进行预测。结果表明，多波长透射光谱的归一化预处理能够有效消除浓度影响，并保留完整的原始光谱信息；通过方差分析法得到特征波长区间为200～300 nm波段，在此区间内提取的五种细菌的归一化光谱趋势图的特征值分别为：200 nm处吸光度值为0.006 5，0.005 1，0.007 5，0.007 5和0.008 5，200～245 nm波段的斜率值为－62.45，－35.94，－81.30，－82.67和－103.49，250～275 nm波段处的斜率值为－15.48，－14.82，－20.91，－13.92和－26.21，280～300 nm波段处的斜率值为－29.96，－24.62，－33.71，－36.09和－30.88。对样本提取特征值并随机划分训练集和测试集，支持向量机选择惩罚因子模型以及线性核函数，通过寻优算法确定最佳的惩罚因子参数 c 和核函数参数 g，对测试集样本进行测试，得到细菌种类的识别结果，五种细菌的预测准确率均达到100.0％。综上所述，水体致病菌的多波长透射光谱通过合适的数据预处理能够提取出具有明显差异性的光谱特征值，该光谱特征值结合支持向量机能够有效用于不同细菌种类的识别，该方法为水体细菌快速识别和实时在线监测提供了重要的技术支持。

关键词 多波长透射光谱；细菌；特征提取；支持向量机；分类识别
中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)09-2940-05

引 言

近年来多波长透射光谱因具有丰富的特征光谱信息成为研究水体致病菌的重要工具。不少学者结合细菌的多波长透射光谱建立光谱解析模型，研究了细菌大小、浓度和化学组分等特征信息的获取方法[1-3]。目前虽然在水体致病菌的多波长透射光谱识别方面已经开展了一定的研究工作[4-5]，但由于不同细菌微生物的光谱相似性较高，且光谱会随细菌微生物所处环境条件的变化而变化，比如浓度、生长阶段等，这些因素大大增加了细菌微生物的识别难度。分析不同细菌多波长透射光谱的特征差异性，可以更好地实现基于多波长透射光谱法的水体致病菌的识别。

在目标分析物的光谱特征提取和光谱识别方法研究方面，高斌等以移动平滑算法处理光谱数据，通过“组合放大”提取光谱特征并基于BP神经网络完成对不同动物血液的荧光光谱识别[6]；宫鹏等研究了高光谱数据处理的一系列方法，结合神经元网络算法实现不同针叶树种的光谱差异分析与识别[7]；张正勇等将紫外可见光谱与化学计量法相结合提取光谱特征，进行白酒年份的鉴别[8]。鉴于此，本文以肺炎

收稿日期：2020-09-10，**修订日期**：2021-01-16
基金项目：国家自然科学基金项目(61875254，61705237，61805254)，安徽省重点研发计划项目(1804a0802192)资助
作者简介：冯 春，女，1994年生，中国科学院安徽光学精密机械研究所博士研究生 e-mail：cfeng@aiofm.ac.cn
*通讯作者 e-mail：njzhao@aiofm.ac.cn；gfyin@aiofm.ac.cn

克雷伯氏菌、金黄色葡萄球菌、鼠伤寒沙门氏菌、铜绿假单胞菌和大肠埃希氏菌为研究对象，获取细菌在不同状态下的多波长透射光谱，对光谱进行归一化处理得到了细菌光谱的最佳测量范围，通过方差分析法得到光谱变动最显著的特征波长区间，在该区间提取 200 nm 处的吸光度值及短波段的斜率值等光谱特征值，结合支持向量机对不同细菌种类进行识别，为水体细菌快速识别和检测提供技术支持。

1　实验部分

1.1　试剂和仪器

肺炎克雷伯氏菌(*K. pneumoniae*)、大肠杆菌(*E. coli*)、鼠伤寒沙门氏菌(*S. typhi*)、金黄色葡萄球菌(*S. aureus*)和铜绿假单胞菌(*P. aeruginosa*)5 种水体常见致病性细菌微生物菌种均购于中国工业微生物菌种保藏管理中心(China Center of Industrial Culture Collection，CICC)；牛肉膏蛋白胨培养基(主要成分及其质量分数，牛肉膏：0.3%，氯化钠：0.5%，蛋白胨：0.5%；pH：7.2)；去离子水等。

紫外可见分光光度计(UV2550，日本岛津)，高速冷冻离心机(H-1650R 型，江东)，压力蒸汽灭菌锅(YX-280D 型，上海华泰)，超净工作台(SW-CJ-ID 型，苏州安泰)，组合式光照振荡培养箱(MQP-B3G 型，上海旻泉)等。

1.2　光谱测量

将液体培养基及所用器皿在 121 ℃下灭菌 20 min 后，在超净台上用接种环挑取斜面固体培养基中的一个细菌菌落，接种到液体培养基中，将接种后的细菌培养液放入培养箱，在温度为 35 ℃，转速为 120 r·min^{-1}条件下进行培养。培养到特定生长阶段，取适量细菌培养液于离心管中，在 12 000 r·min^{-1}的转速下离心 5 min，倒出上清液；再向离心管中倒入去离子水，同样转速下对细菌离心洗涤三次，将离心洗涤后的细菌再次分散在去离子水中，并对该细菌悬浮液进行稀释，获得不同浓度的细菌悬浮液用于细菌多波长透射光谱的测量。

取摇匀后的细菌悬浮液 3.5 mL 加于石英比色皿中进行多波长透射光谱测量，以去离子水为参比扣除背景，消除杂散光。多波长透射光谱测量范围为 200～900 nm，采样间隔为 1 nm。

1.3　光谱数据预处理

消除细菌浓度对光谱的影响，需要对细菌的多波长透射光谱进行归一化处理，根据胡玉霞的研究，总和归一化的光密度谱的标准偏差值最小，该归一化方法得到的细菌浓度反演的结果准确性和稳定性最好[9]。

$$\tau = \tau_i / \mathrm{sum}(\tau_i)$$

其中 τ 表示经总和归一化后的光谱数据，τ_i 表示原光谱数据第 i 个波长点对应的光密度值，i 从 200 到 900。

1.4　光谱特征区间和特征值提取

数据量很大的情况下，需要进行一定的特征提取，或者有些特征之间相互关联，其中一些特征可以用其他特征来表述，利用特征提取来达到问题化简、处理方便的目的。

利用方差分析法计算归一化预处理后的光谱阵在 200～900 nm 区间内各波长的标准偏差，对应标准偏差越大的波长，其光谱变动越显著，给定一阈值来选取用于细菌识别的特征波长区间，在此区间进行光谱特征值的提取。基于 matlab 平台对特征波长区间进行特征值提取，利用 find 函数提取 200 nm 处的光密度值作为第一特征值，利用 polyfit 函数提取 200～245，250～275 和 280～300 nm 波段的斜率值分别作为第二、三、四特征值，得到一个降维后的特征值矩阵。

1.5　支持向量机

支持向量机(support vector machines，SVM)是建立在统计学习理论 VC 维理论和结构风险最小化原理上的机器学习方法[10]，其主要思想是建立一个超平面作为决策曲面，使正反例之间的隔离边缘被最大化。SVM 工具箱种类很多，本研究所用程序采用台大林智仁的 libsvm[11]中的多类模式识别，易于使用且快速有效。在特征提取的基础上，将所有样本随机划分为训练集和测试集，支持向量机选择惩罚因子模型以及线性核函数，通过寻优算法确定最佳的惩罚因子参数 c 和核函数参数 g，再对测试集样本进行测试，得到细菌种类的识别结果。

2　结果与讨论

2.1　数据预处理

以金黄色葡萄球菌为例，将细菌悬浮液进行稀释得到一系列浓度梯度的金黄色葡萄球菌测试样品，并进行多波长透射光谱测量，对测得的光谱进行总和归一化预处理，并将高浓度和低浓度测试样品的归一化光谱进行对比分析，结果如图 1 所示。

由图 1 可以看出，细菌浓度越高，对应的光谱吸光度值越大，高浓度的归一化光谱特征峰更明显但重合度并不高，低浓度的归一化光谱具有很好的重合性，其他四种细菌的光谱图也有类似的规律。当细菌样品浓度高时，吸光粒子间的平均距离减小，受粒子间电荷分布相互作用的影响，摩尔吸收系数发生改变，偏离朗博比尔定律。通过实验确定当细菌的多波长透射光谱最大吸光度值不超过 1.5 a.u. 时，对光谱进行归一化处理可以有效消除浓度对细菌光谱的影响。

以余弦相似度[12]度量归一化处理方法对不同细菌光谱间的影响，余弦相似度越低，说明光谱差异性越大，余弦相似度越高，说明光谱差异性越小。对原始光谱和归一化处理后的光谱差异程度进行对比分析，结果表 1 所示。

由表 1 可知，对光谱进行归一化变换后，各光谱间的相似度值不变，说明归一化处理并不影响不同细菌微生物光谱之间的差异性程度，归一化处理能最大程度保留光谱的原始信息。

2.2　特征提取

选择五种细菌的低浓度样品的光谱曲线各 12 条进行归一化预处理，得到五种细菌的归一化光谱，对细菌的 12 条归一化后的光谱曲线求平均值得到平均值曲线，即为五种细菌各自的归一化光谱图趋势线，结果如图 2 所示。

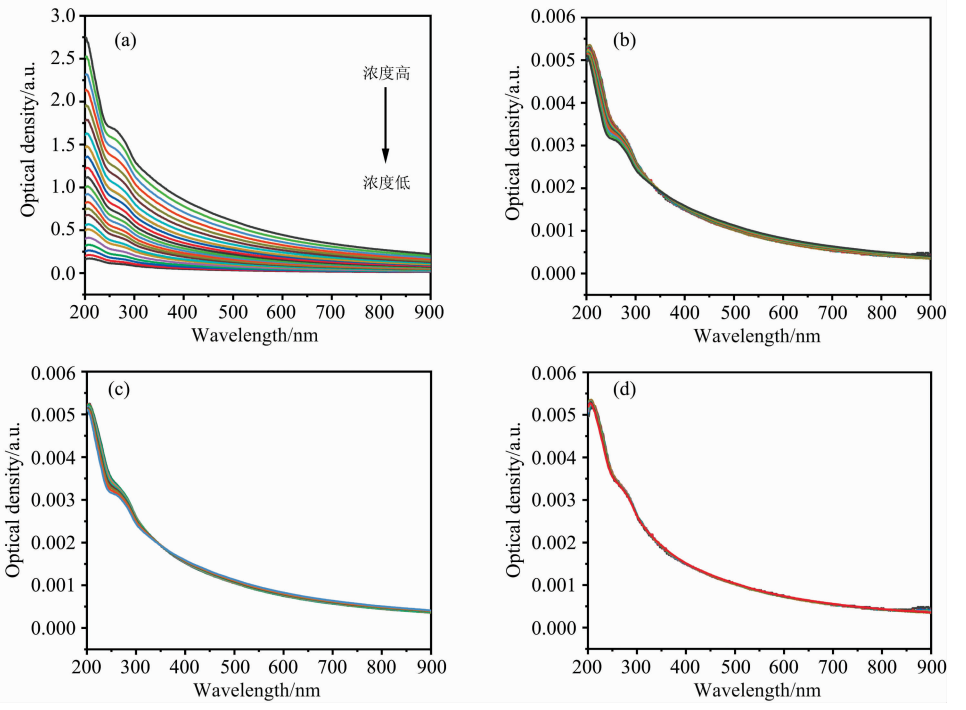

图 1 金黄色葡萄球菌在不同浓度下的多波长透射光谱图及归一化图

(a)：金黄色葡萄球菌不同浓度的光谱图；(b)：各不同浓度光谱图的归一化谱图；
(c)：高浓度光谱图的归一化谱图；(d)：低浓度光谱图的归一化谱图

Fig. 1 Multi-wavelength transmission spectra and normalized images of *S. aureus* at different concentrations

(a): Spectra of *S. aureus* at different concentrations; (b): Normalized spectra of different concentrations;
(c): Normalized spectra at high concentrations; (d): Normalized spectra at low concentrations

表 1 不同细菌微生物光谱之间的余弦相似度

Table 1 Cosine similarity between different microbial spectra of bacteria

1 *E. coli* 2 *K. pneumoniae* 3 *S. aureus* 4 *S. typhi* 5 *P. aeruginosa*
A original B normalized

	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
A	0.999 4	0.964 0	0.997 9	0.988 2	0.967 0	0.997 2	0.985 2	0.946 5	0.915 2	0.995 0
B	0.999 4	0.964 0	0.997 9	0.988 2	0.967 0	0.997 2	0.985 2	0.946 5	0.915 2	0.995 0

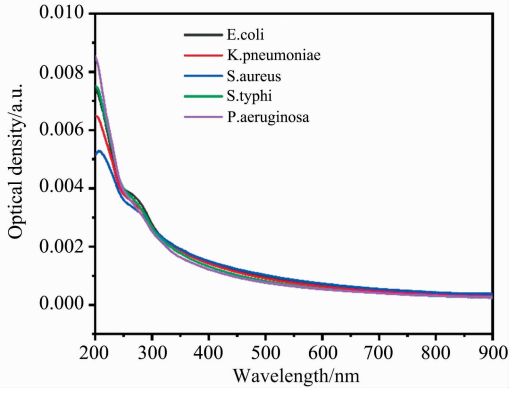

图 2 五种细菌低浓度下的多波长透射光谱归一化趋势图

Fig. 2 Normalized trend graphs of multi-wavelength transmission spectra of five low bacteria concentration samples

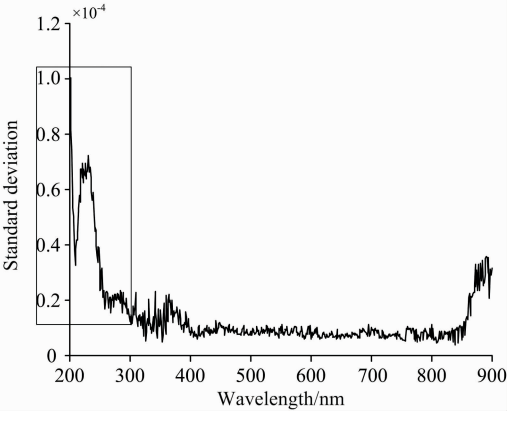

图 3 细菌多波长透射光谱的方差分析图

Fig. 3 Variance analysis diagram of multi-wavelength transmission spectrum of bacteria

利用方差分析法计算归一化预处理后的光谱阵在 200～900 nm 区间内各波长下的标准偏差，标准偏差越大，其光谱变动越显著，通过确定一阈值得到细菌的特征波长区间，根据该区间提取实验样本多波长透射光谱的特征值作为细菌种类识别的光谱数据。方差分析的结果如图 3 所示。

根据图 3 显示，选择 200～300 nm 波段为特征波长区间，在此区间提取不同种类细菌的光谱特征值，选择图 2 中 200 nm 处对应的光密度值，200～245 nm 波段、250～275 nm 波段和 280～300 nm 波段的曲线斜率进行特征值提取，结果如表 2 所示。

表 2　五种细菌微生物的多波长透射光谱归一化图差异性对比

Table 2　The difference of normalized graph of multi-wavelength transmission spectra of five kinds of bacteria

Bacteria	Optical density/a. u. (200 nm)	Slope 1 (200～245 mm)	Slope 2 (250～275 mm)	Slope 3 (280～300 mm)
K. Pneumoniae	0.006 5	−62.45	−15.48	−29.96
S. Aureus	0.005 1	−35.94	−14.82	−24.62
S. Typhi	0.007 4	−81.30	−20.91	−33.71
E. Coli	0.007 5	−82.67	−13.92	−36.09
P. Aeruginosa	0.008 5	−103.49	−26.21	−30.88

从表 2 可以看出五种细菌多波长透射光谱在 200 nm 处的吸光度值及 200～245，250～275 和 280～300 nm 波段的斜率具有差异性，五种细菌在 200 nm 处的吸光度值分别为 0.006 5，0.005 1，0.007 4，0.007 5 和 0.008 5，在 200～245 nm 波段处的斜率值为−62.45，−35.94，−81.30，−82.67 和−103.49，250～275 nm 波段处的斜率值为−15.48，−14.82，−20.91，−13.92 和−26.21，280～300 nm 波段处的斜率值为−29.96，−24.62，−33.71，−36.09 和−30.88。

2.4　样本预测

选取五种细菌 96 个样本，一部分作为训练集(49)，一部分作为测试集(47)，五种细菌的标签分别是 1，2，3，4 和 5。对样本数据进行归一化预处理，分别选取 200 nm 处的吸光度值作为第一特征值，200～245 nm 波段、250～275 nm 波段和 280～300 nm 波段的斜率值作为第二、第三、第四特征值，利用训练集得出的模型对测试集进行预测，得到五种细菌的预测准确率如表 3 所示。

表 3　支持向量机对五种细菌的识别率

Table 3　The recognition rates of five kinds of bacteria by libsvm

Bacteria	Train set	Predict set	recognition rate/%
K. Pneumoniae	9	9	100.0
S. Aureus	8	7	100.0
S. Typhi	10	10	100.0
E. Coli	13	13	100.0
P. Aeruginosa	9	8	100.0
—	49	47	100.0

由表 3 可以看出，五种细菌预测集的预测准确度均达到 100.0%，说明通过归一化数据预处理得到的细菌多波长透射光谱特征值(200 nm 处的吸光度值及 200～245，250～275 和 280～300 nm 波段的斜率值)结合支持向量机(libsvm)可以快速对不同种类的细菌进行识别。将细菌的多波长透射光谱数据直接作为识别模型的输入，训练模型时需处理上百上千的数据，而对光谱数据进行预处理并提取特征值后，模型需处理的数据量不仅大量减少，避免了数据冗余影响模型的识别精度，而且提取的特征值可以最大限度体现不同种类细菌光谱的差异，有助于快速准确识别不同种类的细菌。

3　结　论

对水体细菌进行多波长透射光谱测量，细菌样品的浓度满足最大吸光度值不超过 1.5 a. u. 的情况下，对细菌的多波长透射光谱进行归一化预处理可以有效消除细菌浓度的影响，且保留原始光谱的完整信息。经归一化预处理之后，利用方差分析法得到特征波长区间在 200～300 nm 波段，在该区间提取 200 nm 处的吸光度值及 200～245 nm 波段、250～275 nm 波段、280～300 nm 波段处的斜率特征值，结合支持向量机建立识别模型，能够很好地用于不同细菌种类的识别，对肺炎克雷伯氏菌、金黄色葡萄球菌、鼠伤寒沙门氏菌、铜绿假单胞菌和大肠埃希氏菌的预测准确度可达 100.0%。基于多波长透射光谱技术的特征值提取结合支持向量机在水体细菌微生物快速检测和识别应用方面具有很大潜力。

References

[1] Alupoaei C E, Olivares J A, Garcia-Rubio L H, et al. Biosensors and Bioelectronics, 2003, 19(8): 893.

[2] HU Yu-xia, ZHAO Nan-jing, GAN Ting-ting, et al(胡玉霞, 赵南京, 甘婷婷, 等). China Laser(中国激光), 2018, 45(2): 274.

[3] Alupoaei C E, Garcia-Rubio L H. Biotechnology and Bioengineering, 2004, 86: 163.

[4] Smith J M, Huffman D E, Acosta D, et al. Journal of Biomedical Optics, 2012, 17(10): 1.

[5] Huffman D E, Serebrennikova Y M, Smith J M, et al. Journal of Spectroscopy, 2016, 2016: 5436821(http://dx.doi.org/10.1155/

2016/5368821).
[6] GAO Bin, ZHAO Peng-fei, LU Yu-xin, et al(高 斌, 赵鹏飞, 卢昱欣, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(10): 3136.
[7] GONG Peng, PU Rui-liang, YU Bin, et al(宫 鹏, 浦瑞良, 郁 彬, 等). Journal of Remote Sensing(遥感光学), 1998, (3): 211.
[8] ZHANG Zheng-yong, SONG Chao, SHA Min, et al(张正勇, 宋 超, 沙 敏, 等). Brewing Technology(酿酒科技), 2016, (11): 20.
[9] HU Yu-xia, GAN Ting-ting, ZHAO Nan-jing, et al(胡玉霞, 甘婷婷, 赵南京, 等). Acta Optica Sinica(光学学报), 2018, 38(4): 362.
[10] DING Shi-fei, QI Bing-juan, TAN Hong-yan, et al(丁世飞, 齐丙娟, 谭红艳, 等). Journal of Electronic Science and Technology(电子科技大学学报), 2011, 40(1): 2.
[11] Chang Chih-Chung, Lin Chih-Jen. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.
[12] Liu Donghai, Chen Xiaohong, Peng Dan. International Journal of Intelligent Systems, 2019, 34(7).

Study on Multi-Wavelength Transmission Spectral Feature Extraction Combined With Support Vector Machine for Bacteria Identification

FENG Chun[1,2,3], ZHAO Nan-jing[1,3*], YIN Gao-fang[1,3*], GAN Ting-ting[1,3], CHEN Xiao-wei[1,2,3], CHEN Min[1,2,3], HUA Hui[1,2,3], DUAN Jing-bo[1,3], LIU Jian-guo[1,3]

1. Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China
2. University of Science and Technology of China, Hefei 230026, China
3. Key Laboratory of Optical Monitoring Technology for Environment, Anhui Province, Hefei 230031, China

Abstract The realization of rapid identification of pathogenic bacteria has important practical significance for preventing large-scale disease outbreaks caused by microbial pollution in water bodies. Conventional bacterial detection methods such as biochemical identification and nucleic acid detection have the characteristics of time-consuming and precise experimental equipment, which are insufficient for the rapid and real-time online monitoring of bacteria. Since the multi-wavelength transmission spectrum of bacteria contains abundant characteristic information, and this spectral detection technology has the advantages of fast, simple, non-contact, and non-polluting, it has become a hot spot in bacterial detection research in recent years. This article takes Klebsiella pneumoniae, Staphylococcus aureus, Salmonella typhimurium, Pseudomonas aeruginosa and Escherichia coli as research objects. The characteristic wavelength range with the most significant spectral change is obtained by normalization and the analysis of variance method, and the characteristic spectral values such as the absorbance value at 200nm and the slope value of the short waveband are extracted from this range, and the support vector machine is used to predict different types of bacteria. The results show that the normalization of the multi-wavelength transmission spectrum can effectively eliminate the concentration effect and retain the complete original spectral information. The characteristic wavelength range of 200～300 nm is obtained by analysis of variance. The characteristic values of the normalized spectral trend graphs of the five bacteria extracted in this interval are: The absorbance values at 200 nm are 0.006 5, 0.005 1, 0.007 5, 0.007 5, and 0.008 5. The slope values at the 200～245 nm band are －62.45, －35.94, －81.30, －82.67, and －103.49, and the slope values at the 250～275 nm band are －15.48, －14.82, －20.91, －13.92 and －26.21, the slope values at the 280～300 nm band are －29.96, －24.62, －33.71, －36.09 and －30.88, respectively. Feature values were extracted from the samples and randomly divided into a training sets and test sets. The penalty factor model and the linear kernel function were selected for SVM, the best penalty factor parameter c and kernel function parameter g were determined through the optimization algorithm. The prediction accuracy rates of the five species of bacteria all reach 100.0%. In summary, theobvious spectral characteristic values of the multi-wavelength transmission spectrum of bacteriacan be extracted through proper data preprocessing. The spectral feature value combined with the support vector machine can be effectively used for the identification of different bacterial species. This method provides important technical support for rapid identification and real-time online monitoring of water bacteria.

Keywords Multiwavelength transmission spectrum; Bacterias; Feature extraction; Support vector machine; Classification and identification

* Corresponding authors

(Received Sep. 10, 2020; accepted Jan. 16, 2021)