文章编号:1001-5078(2009)11-1153-05 ·红外技术·

基于可见-近红外光谱技术与BP-ANN算法的污水类型鉴定

刘志霄[1],梁 亮[2],俞晓莹[3]
(1. 吉首大学生物资源与环境科学学院,湖南 吉首 416000;2. 中南大学信息物理工程学院,湖南 长沙 410083;
3. 长沙理工大学土木与建筑学院,湖南 长沙 410076)

摘 要:提出了一种基于可见-近红外光谱技术与BP人工神经网络(BP-ANN)算法快速进行污水类型鉴定的新方法。以FieldSpec ®3地物光谱仪采集了4种污水样品的光谱数据,共168份,随机将其分成校正集(132份)和检验集(36份)。分别采取全波段(400~2450 nm)与择取波段(400~1800 nm)两种方法建立模型进行分析。光谱经S. Golay平滑和标准归一化(SNV)处理后,以主成分分析法(PCA)降维。将降维所得的前9个主成分数据作为BP-ANN的输入变量,污水类型作为输出变量,建立3层BP-ANN鉴别模型。利用36个未知样对模型进行检验。结果表明:两类模型预测准确率均高达100%,且择取波段模型比全波段模型具有更高的预测精度。说明利用可见-近红外技术结合BP-ANN算法进行污水类型的快速、无污染鉴定是可行的,且波段筛选是优化模型的有效方法之一。

关键词:可见-近红外光谱;污水;BP-神经网络;鉴定
中图分类号:S123 **文献标识码**:A

Identification of the types of waste water based on visible/near-infrared spectroscopy and BP-ANN algorithm

LIU Zhi-xiao[1], LIANG Liang[2], YU Xiao-ying[3]
(1. College of Biology and Environmental Sciences, Jishou University, Jishou 416000, China;
2. School of Info- Physics and Geomatics Engineering, Central South University, Changsha 410083, China;
3. Changsha university of science and technology, Changsha 410000, China)

Abstract: A rapidly and pollution-free method was developed to identify the types of waste water by visible/near-infrared spectroscopy and back-propagation artificial neural network (BP-ANN) algorithm. The spectra data of the total 168 samples were obtained by a FieldSpec ®3 spectrometer. All the samples were divided randomly into two groups, one with the 132 samples used as the calibrated set, and the other with the 36 samples as the validated set, and subsequently were analyzed with the whole wave band(400~2450 nm) and the selection wave band(400~1800 nm) models, respectively. The spectra data were pretreated by the methods of S. Golay Smoothing and Standard Normal Variable (SNV), and the pretreated spectra data were analyzed with Principal Component Analysis (PCA). The anterior 9 principal components computed by PCA were used as the input variables of BP-ANN model which included one hidden layer, while the values of the types of waste water used as the output variables, and consequently the three layers BP-ANN identification model was built. The 36 unknown samples in the validated set were predicted by the ANN-BP model. The results showed that the recognition rate was 100% in such both models, and the accuracy of selection wave band model was higher than that of the whole wave band model. We suggested that it was feasible to discriminate the types of waste water used by visible / near-infrared spectroscopy and BP-ANN algorithm as a rapid and pollution-free way, and the wave band selection was a validated way to improve the precision of the identification model.

Key words: visual/near-infrared spectra; waste water; BP-ANN; identification

---

基金项目:国家自然科学基金项目(No. 30570279);中南林业科技大学林业遥感信息工程研究中心开放性研究基金项目(No. RS2008k03);中南大学拔尖博士研究生学位论文创新项目(No. 1960-71131100007);优秀博士论文扶持项目(No. 2008yb024)资助。
作者简介:刘志霄(1965-),男,教授,主要从事生态学研究。E-mail:zxliu1965@163.com
收稿日期:2009-05-16;**修订日期**:2009-08-10

1 引 言

准确获取水污染定性、定量信息是水体污染控制与净化处理的基础,但由于生活、生产以及意外事故均可导致水体污染,故水污染的类型很多,而根据污水类型不同,进行定量分析与净化处理所用方法也不一样,因此在进行方法选择之前,往往还需要实时、快速、准确鉴定污水的类型,为污水后续处理提供基础信息。然而,传统的化学鉴定方法周期长,需要消耗试剂,并产生二次污染,因而有必要探索一种快速、无污染、低成本的分析方法,以实现污水类型的准确鉴定[1]。

可见 - 近红外反射光谱技术因其快速、高效以及无损的特点,已被广泛地应用于石油化工、探矿、制药以及纺织等领域[2]。近年来,可见 - 近红外光谱结合模式识别技术,进一步在农副产品、中药材以及食品的分类鉴定方面获得了成功的应用[3-8]。在对污水的研究方面,一些研究者也已利用这一技术在 BOD_5,COD、总磷以及 pH 值的快速测量方面进行过探索[9-11]。但目前尚未有利用这一技术进行污水类型鉴定的报道。本文采用可见 - 近红外反射光谱技术对污水类型进行分析,为污水类型的快速、无污染、低成本鉴定提供新的方法。

2 实验部分

2.1 实验仪器

FieldSpec ®3 地物光谱仪一台,1000 mL 烧杯若干。FieldSpec ®3 地物光谱仪为美国 ASD(analytical spectral device)公司产品,波长范围为 350 ~ 2500 nm,采样间隔为 1.4 nm,光谱分辨率为 3 nm。

2.2 样本来源与数据采集

污水样本分别采自株洲、湘潭、吉首市各污水处理厂与各厂家排污口,共 168 份。其中生活污水、造酒厂废水样本各 48 份,印染废水、制药厂废水各 36 份。在每类型样本中随机抽取 9 份(共 36 份)作为预测集,其余 132 份作为训练集。样本盛放于烧杯中进行光谱扫描。光谱仪探头置于样本正上方,下部距样本 20 mm;光源与水平面保持 45°角,距样本 500 mm。每一样本重复测量 36 次后取均值。光谱数据在 ASD View Spec Pro 中以 ASCII 码形式导出,再导入 Unscramble 9.7 与 DPS 9.50 中进行处理。

2.3 光谱预处理

由于光谱仪在 400 nm 以前与 2450 nm 以后噪声较大,因此本研究选用 400 ~ 2450 nm 波段作为有效光谱数据进行分析。将有效光谱经 S. Golay 平滑去噪后,采用标准归一化(SNV)进行校正。

2.4 数据降维与波段选择

如果将光谱数据作为变量直接导入进行建模,不但会因变量太多而增加建模难度,而且会引入噪声而导致模型的预测精度降低。为了避免这一问题,本文采用主成分分析(PCA)以实现光谱数据的降维。

2.5 人工神经网络模型

在分类鉴定中,人工神经网络是一种重要的模式识别方法,其中多层误差反向传播神经网络方法(back-propagation,BP)应用尤广,具有很强的非线性建模能力,适合解决复杂的映射问题[12-13]。因此,本研究将降维后的数据导入 DPS 软件[14],采用 BP - ANN 算法建立不同污水类型的鉴别模型。

3 结果与讨论

3.1 污水样本的可见 - 近红外反射光谱

图 1 为 4 种污水样部分可见 - 近红外漫反射光谱曲线。从图中可看出,光谱在 807 nm 与 1070 nm 附近有明显的反射峰,而在其他波段严重吸收,且 1800 nm 后变化很不规则,形成许多毛刺,难以确定是样本的特征信息还是干扰信息。因此,在下一步的研究中,将利用全波段(400 ~ 2450 nm)与择取波段(400 ~ 1800 nm)两种方法进行分析,以确定 1800 nm 以后的波段的存留。分析时,先将光谱数据转化为 ASCII 码,在 Unscramble 9.7 中完成预处理后进行 PCA 分析。

3.2 PCA 分析与主因子提取

训练集样本经 PCA 降维分析后,分别以前 3 个主成分 PC_1,PC_2,PC_3 作为 x,y,z 坐标,建立各样本的三维得分图(如图 2 所示),以表征各样本在该三维空间中的分布情况。由于两种方法前 3 个主成分对光谱矩阵的累积方差贡献分别达 85.94%(全波段)与 96.37%(择取波段),因此,样本在三维空间的分布可大体反应其在超维空间的分布特征,表征出不同类型污水的聚类结果。从图 2 中可看出,各类污水样本具有良好的聚类趋势,可进行初步的判别分析,且可看出择取波段(b)的聚类效果比全波段(a)更好。但要取得更精确的分析结果,还需要建立鉴别能力更强的模型。

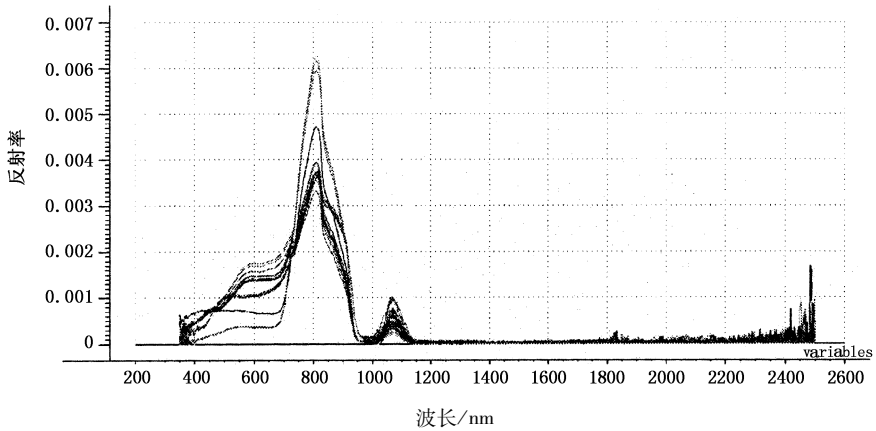
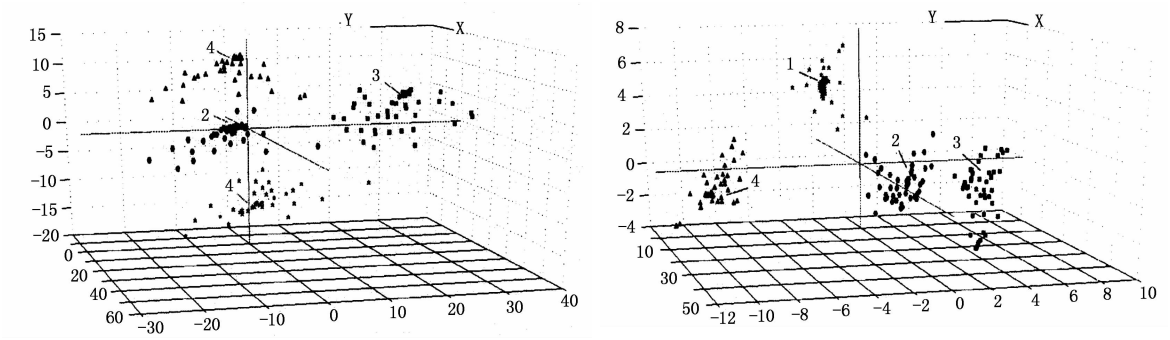


图1 4种污水样本的可见－近红外反射光谱

Fig. 1 the visual/near-infrared reflected spectra of 4 types of waste water



(a) (b)

图2 全波段模型(a)与择取波段模型(b)前3个主成分的得分聚类

Fig. 2 score cluster plot using top three principal components (PCs) for whole wave band model (a) and selection wave band model (b)

注:1－印染废水;2－生活污水;3－造酒厂废水;4－制药厂废水

Note:1－Printing and dyeing waste water;2－Domestic waste water;3－Winemaking waste water;4－Pharmacy waste water

在建立模型过程中,如果选取的主成分过少,将会因不充分拟合而导致模型预测的准确度降低;而若选用的主成分过多,则会产生过拟合现象而导致模型预测精度下降。本文通过交互验证以确定最佳主成分数,即在累积方差贡献(累积可信度)变化不大的情况下选取较少的主成分数。将全光谱与择取光谱分别进行PCA分析后,所得前12个主成分的累积可信度如图3所示。由图可知,两种方法前9个主成分的累积可信度均在93%以上,已提取出了绝大部分的特征信息,意味着若将其作为BP－ANN模型的输入变量,则该模型的变量中已包含了光谱数据绝大部分有价值的信息。

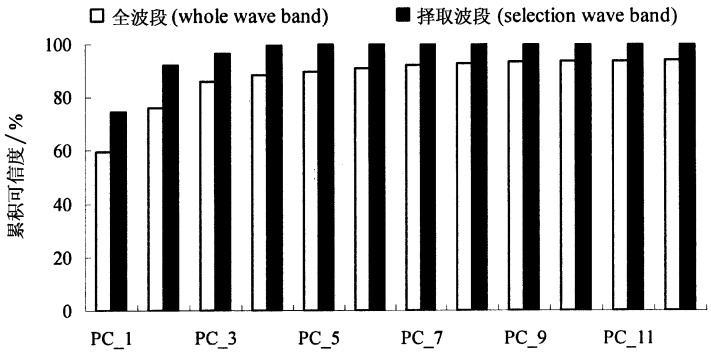


图3 两种模型训练集前12个主成分的累积可信度

Fig. 3 accumulative reliabilities plot of the top 12 principal components of two kinds of PCA models

3.3 BP – ANN 建模与分析

利用训练集中的 132 个样本,以 PCA 降维得到的前 9 个主成分作为 BP 神经网络的输入变量,建立 PCA – BP 神经网络预测模型。建模分析时,将印染废水、生活污水、造酒厂废水与制药厂废水分别赋值为 1.0000,2.0000,3.0000 与 4.0000。BP 网络各层间采用 Sigmoid 激励函数,其中 Sigmoid 参数取 0.9,最小训练速度设为 0.1,动态参数取 0.6,允许误差设为 0.0001,最大迭代次数设为 3000 次。建模过程中,通过调节隐含层的节点数目反复验证以优化 BP – ANN 的结构,得到最佳的 BP 网络结构为 9 –6 –4 三层。利用所建立的 BP – ANN 模型对预测集的 36 个未知样进行预测,结果表明两类模型对所有未知样预测的正确率均达 100%(如表 1 所示)。

表 1 两种模型对 36 个未知样本的预测结果

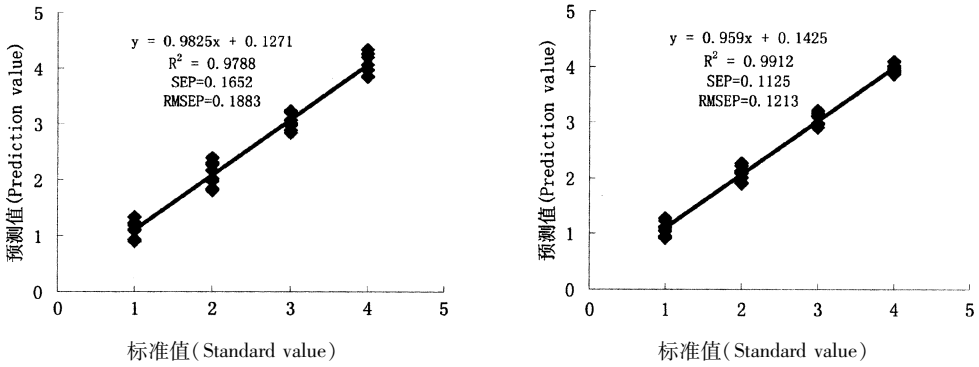
Tab. 1 prediction results for 36 unknown samples by two kinds of BP-ANN models

样本号 Sample No.	标准值 Standard value	预测值(全波段) Predicted value(Whole wave band)	预测值(择取波段) Predicted value(selection wave band)
(1)	1.0000	1.1063	1.1159
(2)	1.0000	1.2263	1.0343
(3)	1.0000	0.8931	1.1116
(4)	1.0000	1.2054	0.9135
(5)	1.0000	1.0840	1.2120
(6)	1.0000	1.3335	0.9432
(7)	1.0000	0.9221	1.0985
(8)	1.0000	1.2054	1.2651
(9)	1.0000	1.1852	1.0591
(10)	2.0000	2.2670	2.1027
(11)	2.0000	1.8387	2.0659
(12)	2.0000	2.1571	2.2133
(13)	2.0000	2.3851	1.9095
(14)	2.0000	1.9697	1.8927
(15)	2.0000	1.7986	2.0865
(16)	2.0000	2.0215	2.2567
(17)	2.0000	2.2932	2.1235
(18)	2.0000	1.9752	1.9954
(19)	3.0000	2.8321	2.9761
(20)	3.0000	3.2061	3.2041
(21)	3.0000	2.9816	2.9531
(22)	3.0000	2.9733	3.1556
(23)	3.0000	3.0552	2.9735
(24)	3.0000	2.8779	3.1057
(25)	3.0000	3.2254	3.0879
(26)	3.0000	3.0158	2.8974
(27)	3.0000	3.2247	3.1021
(28)	4.0000	4.2512	3.9281
(29)	4.0000	4.0592	3.8596
(30)	4.0000	4.3282	4.0876
(31)	4.0000	3.8350	3.9162
(32)	4.0000	4.0671	3.9321
(33)	4.0000	4.1873	3.9676
(34)	4.0000	3.9712	3.9952
(35)	4.0000	3.8545	4.0068
(36)	4.0000	4.1871	3.8945

注:(1) – (9):印染废水;(10) – (18):生活污水;(19) – (27):造酒厂废水;(28) – (36):制药厂废水;

Note:(1) – (9):Printing and dyeing waste water;(10) – (18):Domestic waste water;(19) – (27):Winemaking waste water;(28) – (36):Pharmacy waste water

为筛选出最佳建模方法,分析了两种方法对36个未知样的预测结果。两类模型对预测集的拟合结果与标准值之间的回归关系如图4所示。由图可知,两回归方程的斜率都接近于1,说明两类模型均有良好的预测结果。但择取波段模型各样本点在回归线附近更集中,其预测集决定系数(R^2 =0.9912)比全波段模型(R^2 =0.9788)要高,而预测标准误差(SEP=0.1125)与预测误差均方根(RMSEP=0.1213)均比全波段模型(SEP=0.1652;RMSEP=0.1883)低,说明择取波段模型具有更好的预测效果。可知样本的可见－近红外光谱中,1800 nm后的信息主要是噪声,应该舍去,而择取波段模型是优选方法。



(a) (b)

图4 全波段模型(a)与择取波段模型(b)对未知样本预测值与标准值之间的关系

Fig.4 standard value versus predicted value by whole wave band model (a) and selection wave band model (b) in validation set

4 结 论

对4种不同类型的污水的分析结果表明,采用可见－近红外光谱技术结合BP－ANN算法进行污水类型的鉴别是可行的,从而为污水类型的快速、无污染、低成本鉴别提供了一种新的方法。比较分析表明,利用择取波段所建立的模型比全波段模型具有更高的预测精度,说明波段删选是进行模型优化的有效手段。

参考文献:

[1] 何金成,杨祥龙,王立人,等. 近红外光谱法测定废水化学需氧量[J]. 浙江大学学报:工学版,2007,47(5):752－755,783.

[2] 褚小立,袁洪福,陆婉珍. 近红外光谱仪国内外现状与展望[J]. 分析仪器,2006(2):1－10.

[3] 梁亮,杨敏华,刘志霄,等. 杂交稻种品系与真伪的可见－近红外光谱鉴别[J]. 激光与红外,2009,39(4):407－410.

[4] 韩亮亮,毛培胜,王新国,等. 近红外光谱技术在燕麦种子活力测定中的应用研究[J]. 红外与毫米波学报,2008,27(2):86－90.

[5] 刘燕德,罗吉,陈兴苗. 可见/近红外光谱的南丰蜜桔可溶性固形物含量定量分析[J]. 红外与毫米波学报,2008,27(2):119－122.

[6] 刘沭华,张学工,孙素琴. 中药材产地的近红外光谱自动鉴别和特征谱段选择[J]. 科学通报,2005,50(4):393－398.

[7] 黄敏,何勇,岑海燕,等. 应用可见－近红外光谱技术快速无损鉴别婴幼儿奶粉品种[J]. 光谱学与光谱分析,2007,27(5):916－919.

[8] 裘正军,陆江锋,毛静渊,等. 基于可见－近红外光谱的可乐品牌鉴别方法研究[J]. 光谱学与光谱分析,2007,27(8):1543－1546.

[9] 刘宏欣,张军,王伯光,等. 水质监测中总磷无损的近红外光谱分析研究[J]. 分析科学学报,2008,24(6):664－666.

[10] 何金成,杨祥龙,王立人. 近红外光谱透射法测量废水化学需氧量(COD)的光程选择[J]. 红外与毫米波学报,2007,26(4):317－320.

[11] 何金成,杨祥龙,王立人,等. 基于近红外光谱法的废水COD,BOD_5,pH的快速测量[J]. 环境科学学报,2007,12(12):2105－2108.

[12] Haykin S. Neural network-a comprehensive foundation [M]. New York:Macmillan College Publishing Company, 1994:1－44.

[13] 齐小明,张录达,杜晓林,等. PLS－BP法近红外光谱定量分析研究[J]. 光谱学与光谱分析,2003,23(5):870－872.

[14] 唐启义,冯明光. DPS数据处理系统——实验设计、统计分析及数据挖掘[M]. 北京:科学出版社,2007.