

基于多光谱遥感图像信息的水质污染监测研究

朱小明

(河海大学 计算机与信息学院,江苏 南京 211100)

摘要:在国内陆水资源保护与污染治理事业中,水资源污染监测是其中的核心环节。传统水质监测方法存在需要消耗大量人力物力和时空局限性的问题。随着遥感技术和成像光谱仪的发展,以及机器学习相关领域的快速发展,提出一种基于极限学习机的高分图像水质监测方法。对高分一号卫星所检测到的多光谱水域图像数据信息进行处理,并结合部分时间段的水质检测站实测数据对处理后的 WFV 数据进行标记,搭建基于极限学习机模型对水域进行异常检测,定性判断该水域是否污染,通过对太湖兰山嘴和洪泽湖盱眙淮河大桥区域的水质污染判定,结果表明该方法对水质监测具有一定的可行性和有效性。

关键词:水质监测;遥感图像;数据处理;极限学习机

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2018)11-0052-04

doi:10.3969/j.issn.1673-629X.2018.11.012

Research on Water Quality Monitoring Based on Multi-spectral Remote Sensing Imagery

ZHU Xiao-ming

(School of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: In China's inland water resources protection and pollution control business, water pollution monitoring is one of the core links. Traditional water quality monitoring methods have the problem of consuming a great deal of manpower, material resources and space-time limitations. With the development of remote sensing technology and imaging spectrometers, and the rapid development of machine learning related fields, we propose a GF-1 image water quality monitoring method based on extreme learning machine. The processing of multi-spectral water image data detected by GF-1 satellite is carried out, and the processed WFV data is marked by using the measured data of water quality detection station in some time periods, and then the water quality is determined based on extreme learning machine model. The results show that this method is feasible and effective for water quality monitoring of Taihu Lanshanzui and Hongzehu Xuyi Huaihe River Bridge.

Key words: water quality monitoring; remote sensing image; data processing; extreme learning machine

0 引言

河流和水库是目前世界上最重要的淡水资源,在调整支流,农业灌溉,工业和生活用水,全球生态平衡,防洪排水,沟通航运中具有重要的作用,是全球经济可持续发展与人类生存和发展的重要保证。国内区域广袤,河流水库众多,水库数目超过 100 000 座^[1]。但是由于前期大力发展重工业,未处理过渡排放的工业废水,以及民众对环境保护的意识薄弱,未处理过渡排放的生活污水,导致水生态系统被严重破坏,河流湖泊污染严重,部分水域富营养化,浮游藻类大量繁殖,形成

水华等现象,给国家带来了巨大的经济损失^[2]。根据国家环境监测总站数据显示,国内水域污染等级为 IV 类、V 类、劣 V 类严重污染的占总水域的 80% 以上,其中最严重的事例为 2007 年 5、6 月的太湖蓝藻污染事件^[3],造成了巨大的经济损失。因此,水质污染问题急需解决,而水质污染监测^[4]是其中一项重要环节。

通常对水域质量进行质量分析评价是判断水域污染程度的有效方法。传统的水质监测是对水域进行污染分析监控,国内外学者对此进行了大量研究。例如,黄耀明等采用监测点现场水体采样、提取、分析对海港

收稿日期:2017-12-01

修回日期:2018-04-12

网络出版时间:2018-05-28

基金项目:国家自然科学基金(61170200);国家高分辨率对地观测系统重大专项应用(08-Y30B07-9001-13/15-01);江苏省重点研发计划项目基金(BE2015707)

作者简介:朱小明(1993-),男,硕士研究生,研究方向为机器学习和图像处理。

网络出版地址:http://cnki.net/kcms/detail/61.1450.TP.20180525.1610.068.html

水域进行污染监测^[5];黄建清等使用基于无线传感器网络的水体数据提取进行水质监测^[6];CF Iscen 等对 Uluabat 湖的采样数据进行成分分析以判定水质^[7]等。

随着遥感卫星领域的快速发展,遥感图像在各类研究中应用广泛,其中不同的星载遥感图像的光谱段不同,主要有反射红外遥感、热红外遥感、微波遥感三种类型。其具有探测范围大,收集数据速度快,图像信息丰富,动态监测地面信息等优点,因此在城市规划监测、水库变化监测、灾害监测、地面资源环境监测等方面应用广泛^[8]。其中 X Guan 等^[9]采用 SDT 模型对 Simcoe 湖的 TM 卫星遥感图像进行水质研究分析;汪小钦等^[10]采用基于 TM 影像的波段视反射率进行研究,探讨 TM 影像第四波段,第三波段与水体污染关联的研究等。

而后随着机器学习^[11]的发展,国内外学者又将机器学习带入水质监测研究。例如,J Cao 等提出基于 DAGSVM 的水质评估方法对水质进行监测^[12];宓云骕等^[13]建立 LM-BP 神经网络和支持向量机模型对水质有机污染物浓度进行监测等。

基于上述研究,针对现有水质污染监测研究的时空局限性和监测效果不足以及遥感卫星使用年限等问题,提出基于高分一号 WFV(wide field of view)遥感图像对水域进行定性监测,并引入极限学习机方法,搭建基于集成 ELM 模型对水域进行异常监测。

1 监测区概况

文中研究的水域为太湖兰山嘴地区和洪泽湖盱眙淮河大桥地区。太湖是中国第三大淡水湖,位于长三角地区南缘,横跨江、浙两省,北临无锡,南濒湖州,西依宜兴,东进苏州。太湖水域的面积为 2 338 km²,流域面积为 36 500 km²,水域周边居民人数大约为 3 500 万,其生产总值约占全国八分之一。是国内工业最发达,城镇化水平最高的地区之一。太湖水域平均深度为 1.89 m,最大深度约为 4.8 m,平均年出湖径流量约为 75 亿立方米,其中太湖的蓄水量为 44 亿立方米。在周边城市经济发展,生活用水,交通运输中太湖具有很大的作用。但由于近年来周边城市的工业发展过于迅速,忽略了环境污染问题,导致大量未经处理的污染物流入太湖,使太湖每年都会大面积爆发春夏蓝藻,严重影响了周边城市的生活质量和经济发展。2007 年 5、6 月,太湖蓝藻污染爆发,造成无锡全城自来水污染,导致周边城市生活用水严重短缺,造成了大量经济损失和环境污染。因此太湖的水污染治理受到了政府的高度重视,而水污染监测是其中一项重要环节。

洪泽湖是中国第四大淡水湖,位于江苏省西部淮河下游,苏北平原中部西侧,处于淮安宿迁两市境内,

为淮河中下游结合部。洪泽湖水域面积大约为 2 069 km²,是“南水北调”工程东线部分的组成部分,水域周边居民人数大约为 1 000 万。洪泽湖水域为过水性湖泊,水位波动不稳定,因此水域面积的变化也比较大。洪泽湖大部分情况下蓄水位为 12.5 m,水域容积大约为 31 亿立方米。洪泽湖平均水深为 1.5 m,水域容量最大时是 5 m,由于地势较高,洪泽湖湖底高度一般在 10 m 左右,因此洪泽湖也被称为“悬湖”。由于近年来周边城市发展过快,未经处理的污水大肆排放,有研究表明,洪泽湖从 2000 年到 2008 年,其水域质量大部分时间段都处于严重污染状况。由此可见,洪泽湖水域也需要进行相应的水质污染监测。

2 实验数据与图像预处理

2.1 实验数据

由于太湖的兰山嘴和洪泽湖盱眙淮河大桥这两个位置均处于太湖和洪泽湖的上游,水域的污染物大部分都从这两处流入,造成严重污染,因此国家在此处设立了国家水质自动监测站,可以实时获取这两处的水质信息。其中水域遥感图像的数据选择及处理在 Zhu H 等研究^[14]的基础上重新实验及改进。

(1)太湖兰山嘴(31.216°N,119.913°E)。本次实验从中国资源卫星应用中心数据库挑选了由高分一号卫星(GF-1)拍下的 13 幅成像效果比较好的 16 m 分辨率的 WFV 遥感图像,其中这 13 幅图像为每天一幅,其成像时间为 2013 年到 2016 年。从每幅 WFV 图像中选取了位于兰山嘴附近的 10 * 10(100)个像素点的遥感波段值,数据集总共为 1 300 条数据,每条 WFV 数据由蓝色、绿色、红色、近红外四个波段组成。同时根据兰山嘴国家水质自动监测站发布的实测数据,将获取到的 13 幅遥感图像根据水质状况分为Ⅱ类、Ⅲ类、Ⅳ类、Ⅴ类、劣Ⅴ类。根据相关规定,等级为Ⅱ类、Ⅲ类的水质为正常水质,等级为Ⅳ类、Ⅴ类、劣Ⅴ类的水质为污染水质,实验测试中将正常水质数据标记为“+1”,共有 600 条数据,将污染水质数据标记为“-1”,共有 700 条,如表 1 所示。

表 1 太湖兰山嘴 WFV 图像日期与污染等级

水质	图像日期			标签
Ⅱ类	2013-12-24	2013-12-29	2014-01-02	+1
Ⅲ类	2013-12-04	2013-12-12	2014-01-10	+1
Ⅳ类	2013-05-12	2013-10-13	2013-11-18	-1
Ⅴ类	2015-06-06	2015-10-02	2015-10-15	-1
劣Ⅴ类	2015-08-01			-1

(2)洪泽湖盱眙淮河大桥(33.030°N,118.495°

E)。本次实验从中国资源卫星引用中心数据库挑选了由高分一号卫星(GF-1)拍下的 15 幅成像效果比较好的 16 m 分辨率的 WFV 遥感图像,其中这 15 幅 WFV 图像为每天一幅,其成像时间为 2013 年到 2016 年。由于水域快读限制,实验从每幅 WFV 图像中选取了位于洪泽湖盱眙淮河大桥附近的 $5 \times 10(50)$ 个像素点的遥感波段值,数据集总共为 750 条数据,每条 WFV 数据由蓝色、绿色、红色、近红外四个波段组成。同时根据洪泽湖国家水质自动监测站发布的实测数据,将所获取到的 15 幅遥感图像根据水质状况分为Ⅱ类、Ⅲ类、Ⅳ类、Ⅴ类、劣Ⅴ类。根据相关规定,等级为Ⅱ类、Ⅲ类的水质为正常水质,等级为Ⅳ类、Ⅴ类、劣Ⅴ类的水质为污染水质,实验测试中将正常水质数据标记为“+1”,共有 500 条数据,将污染水质数据标记为“-1”,共有 250 条,如表 2 所示。

表 2 洪泽湖盱眙淮河大桥 WFV 图像日期与污染等级

水质	图像日期			标签
Ⅱ类	2013-12-12	2013-12-13	2015-05-25	+1
	2013-08-01	2014-01-10	2014-01-14	
Ⅲ类	2014-05-08	2014-05-09	2015-03-28	+1
	2015-07-30			
Ⅳ类	2013-08-10	2014-01-18	2014-01-22	-1
	2014-04-30	2015-07-13		

2.2 图像预处理

文中图像数据是从中国资源卫星应用中心获取,是由高分一号(GF-1)所拍摄,但由于所拍摄图片干扰辐射较多,所以需要先进行以下矫正:

- (1) 辐射定标。校正传感器本身的误差,使用 EN-VI5.1 软件的 Radiometric Calibration 工具对光谱图像进行相应的辐射定标。
- (2) 大气校正。校正大气散射引起的误差,使用 ENVI5.1 软件的 FLAASH 模型工具对光谱图像进行相应的大气校正。
- (3) 正射校正。校正图像中的地形影响,使用 EN-VI5.1 软件的 RPC Orthorectification 工具进行相应地正射校正。

3 研究方法

3.1 极限学习机

极限学习机(extreme learning machine, ELM)是一种单隐层神经网络算法^[15],广泛应用于回归与分类领域^[16],其相比于其他单隐层前馈神经网络(SLFNs)具有学习速度快,不会陷入局部最优且泛化性能好的优点。

极限学习机的原理是在学习过程中网络随机初始化成输入权重和偏置值,只需设置网络的隐层节点

个数,求解最优输出矩阵,使得误差最小。ELM 网络结构如图 1 所示。

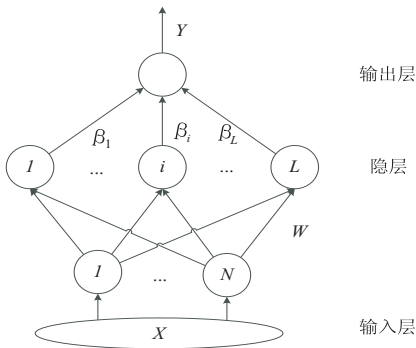


图 1 ELM 网络结构

假设有 N 个任意样本 (X_j, t_j) , 那么对于图 1 中含有 L 个隐层节点的 ELM 网络可以表示为:

$$\sum_{i=1}^L \beta_i g(W_i \cdot X_j + b_i) = Y_j, j = 1, 2, \dots, N \tag{1}$$

其中, $g(x)$ 为激活函数; W_i 为输入层与隐含层权值; b_i 为输入层与隐含层偏置; β_i 为隐含层与输出层的连接权值。

该网络的学习目标是让网络的输出与样本误差最小,表示如下:

$$\sum_{j=1}^N \|Y_j - t_j\| = 0 \tag{2}$$

由此可推出,存在 β_i, W_i, b_i , 使得:

$$\sum_{i=1}^L \beta_i g(W_i \cdot X_j + b_i) = t_j, j = 1, 2, \dots, N \tag{3}$$

使用矩阵表示为:

$$H\beta = T \tag{4}$$

其中, H 为该网络隐层节点的输出; β 为隐含层与输出层的连接权值; T 为样本期望输出。

$$H = \begin{bmatrix} g(W_1 \cdot X_1 + b_1) & \cdots & g(W_L \cdot X_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(W_1 \cdot X_N + b_1) & \cdots & g(W_L \cdot X_N + b_L) \end{bmatrix}_{N \times L} \tag{5}$$

由此可知,当激活函数可微时,由于之前随机的输入权重和偏置固定,可以推出解为:

$$\beta = H^+ T \tag{6}$$

其中, H^+ 为 H 矩阵的广义逆(Moore-Penrose)。

3.2 极限学习机集成

由于单个 ELM 模型的稳定性可能存在不足,故文中选择 N 个 ELM 模型对训练样本进行训练,最后利用极限学习机集成方式对所有数据进行预报,以获得较高的预报精确度和稳定性。

- 基于 ELM 集成的水域质量监测方法的步骤为:
- Step1:根据训练样本训练 N 个极限学习机;
- Step2:使用训练好的 N 个极限学习机对训练样本

进行预测,根据 N 个 ELM 模型预测下的正确率,选取并记录前 k 个模型的正确率 $\alpha_i (i = 1, 2, \cdots, k)$;

Step3:使用挑出的前 k 个模型对测试样本进行预测,记录结果为 $Q_i (i = 1, 2, \cdots, k)$;

Step4:采用加权平均法对 k 个极限学习机进行集成,根据 Step2 的正确率 α_i 确定模型的权值 W_i ,正确率 α_i 越高,权值 W_i 越大,公式如下:

$$W_i = \frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$$

(7)

Step5:根据 Step3 的预测结果 Q_i 和 Step4 的权值 W_i ,计算得到最终的预测结果 Q 。

$$Q = \sum_{i=1}^k W_i Q_i$$

(8)

3.3 实验流程

实验流程如图 2 所示。

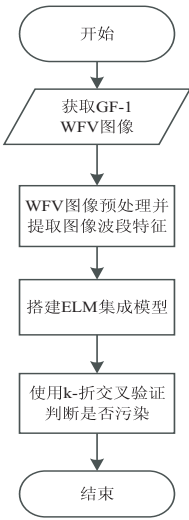


图 2 实验流程

4 结果与分析

文中使用集成极限学习机搭建模型分别对太湖兰山嘴和洪泽湖盱眙淮河大桥进行实验,模型中 N 取值为 15,每条数据预测出的结果如果大于 0 则为 1,小于 0 则为-1。

实验一:从高分一号所获得的 WFV 图像数据中,分别将两个地点的总数据集中随机分为 4 等份,搭建 ELM 模型,使用 k -折交叉验证, $k = 4$,得出实验预测像素点污染准确率,结果如表 3 所示。

表 3 实验一分类准确率 %

水质监测	分类准确率	
	集成 ELM	ELM
太湖兰山嘴	97.7	94.3
洪泽湖盱眙淮河大桥	95.2	92.2

数据,按天分为 13 等份,每天一幅 WFV 图像,搭建 ELM 模型,使用 k -折交叉验证, $k = 13$,得出实验预测准确率。将高分一号所获得的洪泽湖盱眙淮河大桥 WFV 数据,按天分为 15 等份,每天一幅 WFV 图像,搭建 ELM 模型,使用 k -折交叉验证, $k = 15$ 。其中每幅图像中重度污染的像素点数大于 60%,则该图像所在区域当日为重度污染,反之则为正常水质。实验结果如表 4 所示。

表 4 实验二分类准确率 %

水质监测	集成 ELM
太湖兰山嘴	96.3
洪泽湖盱眙淮河大桥	94.8

从表 3 可以看出,水质污染与水质的遥感光谱图像具有一定的相关性,而集成极限学习机模型可以很好地通过该水质的光谱数据对水质污染程度进行区分;从表 4 可以看出,集成 ELM 对水质的预测效果比较好,而其中洪泽湖盱眙淮河大桥比太湖兰山嘴的准确率低,其原因可能是淮河大桥那边的水体由于在城镇附近,干扰物比较多,总体来说实验准确率比较高。

5 结束语

文中提出了一种结合高分一号光谱图像和机器学习极限学习机模型对水质进行监测的新方法,采用该方法对太湖兰山嘴和洪泽湖盱眙淮河大桥进行实验监测。结果表明,该方法能够定性地判别一个地区当日是否严重污染,在训练样本相同的情况下,该模型与使用单个极限学习机判别预测相比,具有更好的准确率和稳定性,与其他机器学习模型如 SVM 模型相比具有更好的速度和泛化性,与传统模型相比具有更好的即时性和操作简便性。因此,该方法可以有效地对水质进行污染监测,具有一定的参考价值。

参考文献:

[1] 马荣华,杨桂山,段洪涛,等. 中国湖泊的数量、面积与空间分布[J]. 中国科学:地球科学,2011,41(3):394-401.

[2] 袁旭音. 中国湖泊污染状况的基本评价[J]. 火山地质与矿产,2000,21(2):128-136.

[3] 李秉柏,施德堂,王志明. 太湖蓝藻暴发的原因及对策建议[J]. 江苏农业科学,2007(6):336-339.

[4] 隋战鹰. 浮游藻类与水质污染监测[J]. 生物学通报,2002,37(8):49.

[5] 黄耀明,沈先荣,罗群,等. 某海港水质监测及其油类污染源分析[J]. 中华航海医学与高气压医学杂志,2011,18(4):236-238.

[6] 黄建清,王卫星,姜晟,等. 基于无线传感器网络的水产养殖水质监测系统开发与试验[J]. 农业工程学报,2013,

实验二:将高分一号所获得的太湖兰山嘴 WFV

- content centric networking under flash crowds[J]. Journal of Network & Computer Applications, 2015, 50: 101–113.
- [7] TANG Xueyan, CHANSON S T. Coordinated en-route web caching[J]. IEEE Transactions on Computers, 2002, 51(6): 595–607.
- [8] 刘外喜, 余顺争, 蔡 君, 等. ICN 中的一种协作缓存机制[J]. 软件学报, 2013, 24(8): 1947–1962.
- [9] LI Zhe, SIMON G. Time-shifted TV in content centric networks; the case for cooperative in-network caching[C]//IEEE international conference on communications. Kyoto, Japan; IEEE, 2011: 1–6.
- [10] JIANG A, BRUCK J. Optimal content placement for en-route web caching[C]//IEEE international symposium on network computing and applications. Cambridge, MA, USA: IEEE, 2003.
- [11] LAOUTARIS N, CHE Hao, STAVRAKAKIS I. The LCD interconnection of LRU caches and its analysis[J]. Performance Evaluation, 2006, 63(7): 609–634.
- [12] LAOUTARIS N, SYNTILA S, STAVRAKAKIS I. Meta algorithms for hierarchical Web caches[C]//IEEE international conference on performance, computing, and communications. Phoenix, AZ, USA: IEEE, 2005: 445–452.
- [13] PSARAS I, WEI K C, PAVLOU G. Probabilistic in-network caching for information-centric networks[C]//Proceedings of the second edition of the ICN workshop on information-centric networking. Helsinki, Finland; ACM, 2012: 55–60.
- [14] BERNARDINI C, SILVERSTON T, FESTOR O. MP C: popularity-based caching strategy for content centric networks[C]//IEEE international conference on communications. Budapest, Hungary: IEEE, 2013: 3619–3623.
- [15] 芮兰兰, 彭 昊, 黄豪球, 等. 基于内容流行度和节点中心度匹配的信息中心网络缓存策略[J]. 电子与信息学报, 2016, 38(2): 325–331.
- [16] CHAI W K, HE Diliang, PSARAS I, et al. Cache “less for more” in information-centric networks (extended version) ☆☆☆[J]. Computer Communications, 2013, 36(7): 758–770.
- [17] MEGIDDO N, MODHA D S. Outperforming LRU with an adaptive replacement cache[J]. Computer, 2004, 37(4): 58–65.
- +++++
- (上接第 55 页)
- 29(4): 183–190.
- [7] ISCEN C F, EMIROGLU Ö, ILHAN S, et al. Application of multivariate statistical techniques in the assessment of surface water quality in Ulubat Lake, Turkey[J]. Environmental Monitoring & Assessment, 2008, 144(1–3): 269–276.
- [8] 喻 欢, 林 波. 遥感技术在湖泊水质监测中的应用[J]. 环境科学与管理, 2007, 32(7): 152–155.
- [9] GUAN Xian. Monitoring Lake Simcoe water quality using Landsat TM images[J]. Corrosion Science, 2009, 51(4): 703–706.
- [10] 汪小钦, 王钦敏, 刘高焕, 等. 水污染遥感监测[J]. 遥感技术与应用, 2002, 17(2): 74–77.
- [11] 闫友彪, 陈元琰. 机器学习的主要策略综述[J]. 计算机应用研究, 2004, 21(7): 4–10.
- [12] CAO Jian, HU Hongsheng, QIAN Suxiang, et al. Research on water quality assessment method based on multi-class support vector machines[C]//International conference on control, automation, robotics and vision. Hanoi, Vietnam: IEEE, 2008: 1661–1665.
- [13] 宓云翥, 王晓萍, 金 鑫. 基于机器学习的水质 COD 预测方法[J]. 浙江大学学报: 工学版, 2008, 42(5): 790–793.
- [14] LI Shijin, ZHU Haichen, CHEN Deqing, et al. Water quality monitoring based on multiple remote sensing imageries[C]//International workshop on earth observation and remote sensing applications. Guangzhou, China: IEEE, 2016: 112–116.
- [15] HUANG Guangbin, ZHU Qinyu, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1–3): 489–501.
- [16] HUANG Guangbin, ZHOU Hongming, DING Xiaojian, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2012, 42(2): 513–529.