J. Lake Sci.(湖泊科学),2015,**27**(1):38-43
http://www.jlakes.org. E-mail:jlakes@niglas.ac.cn

基于支持向量机分类的嘉陵江草街水库甲藻水华预警*

刘朔孺[1,2],杨　敏[1],张方辉[1],张　晟[1**]
(1:重庆市环境科学研究院,重庆 401147)
(2:重庆大学城市建设与环境工程学院,重庆 400045)

摘　要:嘉陵江草街水库自建成后2011－2013年连续3年发生甲藻水华现象,给当地经济发展和生态安全带来影响.根据2011年5月至2013年7月草街水库大坝上、下游8个断面的逐月调查数据,利用支持向量机在处理小样本问题、非线性分类问题和泛化推广方面的优势,构建了基于支持向量机分类的草街水库甲藻水华预警模型.结果表明,利用本月理化数据和本月倪氏拟多甲藻(*Peridiniopsis niei*)密度数据建立的模型,对测试样本取得了80%以上的判别正确率,且对甲藻水华样本的判别正确率为100%.因此,支持向量机作为新兴的机器学习方法,可以为环境管理部门发布水华预警信息提供科学依据,并在环境保护领域具有广阔的应用前景.
关键词:支持向量机;甲藻水华;草街水库;倪氏拟多甲藻

Research on early warning of dinoflagellate bloom in Caojie Reservoir base on support vector machine classification

LIU Shuoru[1,2], YANG Min[1], ZHANG Fanghui[1] & ZHANG Sheng[1]
(1: *Chongqing Academy of Environmental Science*, *Chongqing* 401147, *P. R. China*)
(2: *Faculty of Urban Construction and Environmental Engineering*, *Chongqing University*, *Chongqing* 400045, *P. R. China*)

Abstract: Dinoflagellate bloom consecutively occurred in Caojie Reservoir from 2011 to 2013 and threatened the local economy and ecology. Recently, support vector machine(SVM) was reported to have advantages of only requiring a small amount of samples, high degree of prediction accuracy, and generalization to solve the nonlinear classification problems. In this study, the SVM-based prediction model for dinoflagellate bloom was established by monthly field date collected from May 2011 to July 2013 at 8 transects in Caojie Reservoir. The maximum accuracy excessed 80% by choosing environmental variables data and *Peridiniopsis niei* abundance of current month, and accuracy arrived at 100% for dinoflagellate bloom samples. The results showed that the SVM classification is an effective new way that can be used in monitoring dinoflagellate bloom in Caojie Reservoir and have a promising application prospect for environmental protection.
Keywords: Support vector machine; dinoflagellate bloom; Caojie Reservoir; *Peridiniopsis niei*

近年来,随着我国经济的高速增长,环境生态问题日益突出.而湖泊水库作为居民饮用水重要的水源地,其富营养化已成为影响我国居民生活质量的一个普遍性问题.目前,我国有66%的湖泊、水库处于富营养化水平[1],并且近年来全国不同地区水华的频发更是令人担忧,太湖、巢湖、滇池的蓝藻水华[2-4],三峡水库上游各支流甲藻水华均对当地人民用水安全造成严重影响[5-6],因此采取有效措施防治水华已成为当地环境管理部门的当务之急.

目前国内外对于淡水水华监测预警研究多集中在蓝藻方面,提出了蓝藻水华形成的"四阶段理论"[7],并利用卫星遥感和数学模型成功对蓝藻水华的运动趋势和发生时间进行预测预警[8-11].与大多数蓝藻水华

* 重庆市环境保护局环保科技项目(环科字2012第02号)和重庆市基本科研业务费计划项目(2013cstc-jbky-01604)联合资助.2014－01－02收稿;2014－05－27收修改稿.刘朔孺(1985～),男,博士研究生;E-mail:lsrzggod@163.com.
** 通信作者;E-mail: shengzsts@126.com.

种类不同,甲藻具有垂直迁移特性,其白天趋于在水体表层聚集分布,晚上趋于在水体中随机分布,因此其水华形成和消亡机制与蓝藻具有明显差别[12]. 虽然近些年国内已有关于淡水甲藻水华的报道,但均为水华暴发原因和暴发后的调查研究[5-6,13-16],对于甲藻水华的预警研究还鲜有报道.

Vapnik 等于 1995 年提出的支持向量机(SVM)机器学习方法属于数据驱动模型[17],它不仅克服了水质生态模型对于大量数据样本的需要,还解决了多元统计回归简单线性化的问题. 而与神经网络等传统机器学习方法相比,也不必考虑参数与网络结构的调整,并且模型输出结果易于解释[18]. 目前国内外利用支持向量机进行水华预警也取得了一定进展[18-19],本文利用支持向量机分类方法对嘉陵江草街水库甲藻水华进行预测,以期补充国内淡水甲藻水华预警研究方面的空白,并为及时有效地控制草街水库甲藻水华提供科学依据.

1 材料与方法

1.1 研究区域概况

嘉陵江(29°20′~34°25′N,103°45′~109°0′E)是长江支流中流域面积最大的河流,全长 1119 km. 其中合川市以下至河口段为下游段,大部流经盆地东部平形岭谷地带,属亚热带季风性湿润气候,冬、夏季较长,春、秋季较短,年平均温度 18.2℃,平均降雨量 1126 mm,降雨主要集中在 5－10 月. 草街水库位于嘉陵江江口以上 68 km 处的合川区草街镇,总面积 72.4 km^2,以发电、供水、拦沙为主要功能,该水库于 2010 年建成. 伴随着水库蓄水,大坝下游嘉陵江水位显著下降,而上游水位随之升高,水流减缓,水环境由典型的河流水体转变为类似湖泊的缓流水体. 在水体营养物质浓度不变的条件下,水体流速降低导致泥沙和营养物质的沉淀,为藻类的生长繁殖提供了有利条件,在水库建成后的 2011－2013 年,连续 3 年发生了由倪氏拟多甲藻(*Peridiniopsis niei*)引起的水华,不仅破坏了嘉陵江生态环境,而且威胁着当地居民生活饮用水安全.

1.2 样品采集与分析

本次研究共设置 8 个采样断面,其中金子、玉溪、码头、三江汇合和坝上断面分布在草街水库大坝上游,坝下、梁沱水厂和大溪沟水厂断面分布在草街水库大坝下游,并且每个断面包括左、中、右 3 个采样点(图 1).

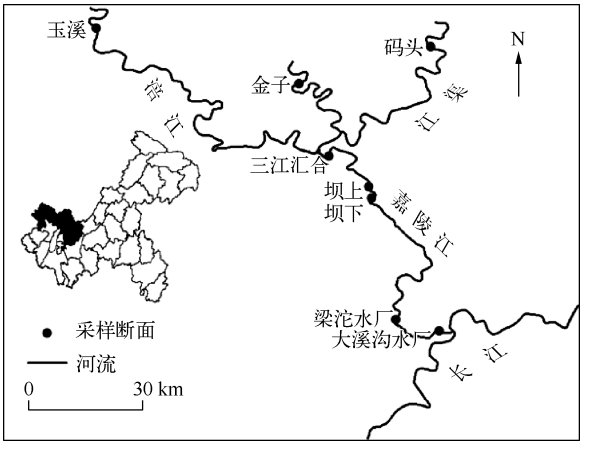

图 1 草街水库采样断面

Fig. 1 Sampling transects in Caojie Reservoir

从 2011 年 5 月至 2013 年 7 月逐月对所有样点进行水样与浮游植物样品的采集,其中流速、水温、溶解氧、浊度、透明度、电导率在野外直接测定. 总氮、硝酸盐离子、亚硝酸盐离子、铵氮离子、总磷、磷酸盐离子、高锰酸盐指数、光照强度和水体叶绿素 a 于实验室根据国家环境保护总局《水和废水监测分析方法》进行测定[20]. 于水面下 0.5 m 处采集 1 L浮游植物定量样品沉淀浓缩计数得到倪氏拟多甲藻密度.

1.3 模型方法

支持向量机作为一种广泛应用的机器学习工具,它既克服了传统方法的大样本要求,还有效地克服了维数灾难及局部极少问题. 模型泛化能力强. 计算简单以及在处理非线性问题时显示的优越性都为其在水质评价与预警研究方面提供了巨大的应用前景.

支持向量机从功能上分为分类与回归两类,本研究中甲藻水华预警模型以其分类功能为基础. 支持向量机分类的基本思想是在样本空间或特征空间构造出最优超平面,使得超平面与不同类样本集之间的距离最大,从而达到最大的泛化能力. 设线性可分样本集为 $T=\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$,$X_i\in R^m$,$Y_i\in\{-1,1\}$,$i=1,2,\cdots,n$,分类面方程为$(\omega\cdot X)+b=0$,分类判别如下:

$$Y_i[(\omega\cdot X_i)+b]-1\geqslant 0 \tag{1}$$

式中,使等号成立的向量称为支持向量. 在 2 类样本线性可分的状况下,求解基于最优超平面的决策函数,可以看成求解二次规划问题. 由解析几何可知类间间隔为 $D = 2/\|\omega\|$,问题可转化为使函数 $\Phi(\omega) = \|\omega\|^2/2$最小化,引入 Lagrange 函数求解这一最优化问题:

$$L = \frac{\|\omega\|^2}{2} - \sum_{i=1}^{n} \alpha_i Y_i [(X_i \cdot \omega) + b] + \sum_{i=1}^{n} \alpha_i \tag{2}$$

其中 $\alpha_i > 0$ 为 Lagrange 乘子,根据 Kuhn-Tucker 条件,这一问题的解必须满足:

$$\alpha_i \{[(X_i \cdot \omega) + b] Y_i - 1\} = 0 \tag{3}$$

一般情况下,大多数样本 α_i将为 0,取值不为 0 的 α_i所对应的样本就是支持向量. 求解上述问题后所得到判别函数为:

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{n} Y_i \alpha_i (X_i \cdot X) + b\} \tag{4}$$

考虑到一些样本不能被超平面正确分类,通过引入惩罚参数(C)和松弛变量(ξ)修正函数,确保模型具有良好的容错性[21].

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} \xi_i \tag{5}$$

对于线性不可分样本,支持向量机借助核函数 K 进行非线性变换,将样本数据映射到高维特征空间中,变为线性问题,再求取最优超平面,然后映射回原空间的非线性分类. 核函数是支持向量机实现空间内积转换运算的函数形式,它不但解决了低维空间的线性不可分,还克服了高维空间的维数灾难,核技巧的应用更使我们避免了因维数增大而导致的巨量计算[22]. 常用的核函数有多项式函数、径向基函数(Radial Basis Function,RBF)、sigmoid 函数等[23].

2 结果与讨论

2.1 模型参数构建

从空间尺度上考虑到草街水库大坝上下游生境变化较大,所以对于大坝上游 5 个断面和下游 3 个断面分别建模. 从时间尺度上考虑分为 2 个类型,一个类型为使用本月水体理化数据预测本月甲藻水华,另一类型为使用本月水体理化数据预测下月甲藻水华,这样共建立 4 个甲藻水华预警模型(表 1).

表 1 预测模型类型

Tab. 1 Prediction model types

	本月理化数据预测本月甲藻水华	本月理化数据预测下月甲藻水华
草街水库大坝上游样点	模型 1	模型 3
草街水库大坝下游样点	模型 2	模型 4

通常认为甲藻水华的发生是多种因素共同作用的结果,它们间的作用机制十分复杂[24],因此只能在众多的环境变量中提取主要的环境因子来建立模型. 本次研究使用 Spearman 相关性分析,将与倪氏拟多甲藻密度显著相关,并且相关系数大于 0.3 的理化变量筛选出来(表 2),然后对筛选出的理化数据进行标准化处理,再进行 PCA 分析,其累积方差贡献率达到 90% 的主成分最终进入模型(表 3).

有关研究表明,当水体中甲藻密度达到 1×10^5 cells/L 时,水面开始明显出现块状褐色分布,到 1×10^6 cells/L 时,用肉眼就能观察到明显的水华现象[15]. 过去 3 年对草街水库浮游植物例行监测发现,主要的甲藻水华优势种为倪氏拟多甲藻,所以本次研究将倪氏拟多甲藻密度 1×10^5 cells/L 设为水华预警的警戒值,即当水体中倪氏拟多甲藻密度达到 1×10^5 cells/L 时就认为有发生甲藻水华的风险,当地环境管理部门需要采取相应的预防措施以减少水华可能带来的问题,此时为甲藻水华样本组,标签为“1”,反之为非甲藻水华样本组,标签为“0”.

在建立模型前,将样本分为2组,2011年与2012年的数据作为训练组,2013年的数据作为检验组,用训练组样本进行模型建立,用检验组样本检验模型的准确率. 通过MATLAB软件对使用不同核函数的判别模型进行比较,最终确定采用误差最小、分类准确率最高的径向基函数作为核函数. 为了确定最优参数,本研究分别使用了网格法、遗传算法和粒子群算法进行参数寻优,并从中选取判别误差最小的参数组合建立甲藻水华预警模型(表4).

表2 Spearman 秩相关性分析结果*

Tab. 2 Spearman correlation analysis of prediction models

指标	模型1	模型2	模型3	模型4
流速	-0.257	-0.178	**-0.399**	**-0.378**
水温	**-0.340**	**-0.367**	**-0.614**	**-0.583**
溶解氧	**0.421**	**0.516**	**0.524**	**0.525**
浊度	**-0.491**	**-0.333**	**-0.449**	-0.253
透明度	**0.452**	0.274	**0.467**	**0.369**
电导率	**0.338**	0.296	**0.433**	**0.353**
总氮	-0.233	-0.125	-0.149	0.108
硝酸盐	-0.099	-0.037	0.023	0.157
亚硝酸盐	0.137	-0.026	0.018	0.064
铵氮	0.284	**0.354**	**0.300**	**0.469**
总磷	**-0.333**	**-0.336**	-0.277	-0.201
可溶性磷酸盐	-0.298	**-0.417**	-0.273	**-0.318**
高锰酸盐指数	**-0.346**	**-0.366**	-0.254	-0.281
叶绿素 a	**0.314**	0.145	0.144	-0.024
光照强度	0.174	0.190	-0.150	-0.185

* 黑体表示相关系数大于0.3.

表3 各模型主成分贡献率(%)

Tab. 3 The contribution of principal components for the models

	模型1	模型2	模型3	模型4
PC1	47.41	37.31	42.75	45.99
PC2	18.77	33.63	20.48	22.30
PC3	10.21	13.04	15.17	11.16
PC4	6.91	7.83	8.40	9.23
PC5	5.74	3.94	6.04	5.72
PC6	4.44	2.57	4.21	2.83
PC7	3.79	1.68	2.96	2.77

表4 模型参数寻优结果

Tab. 4 The results of parameter optimization for the models

	参数寻优方法	惩罚参数 C	核参数 γ	松弛变量 ξ
模型1	粒子群算法	61.5757	1.8692	0.1
模型2	网格法	6.9644	12.1257	0.1
模型3	粒子群算法	18.8545	15.9569	0.1
模型4	遗传算法	5.5599	30.4346	0.1

2.2 预测结果

4个模型对于测试样本总体判别正确率均达到80%,使用本月理化数据和本月倪氏拟多甲藻密度建立

的模型,甲藻水华样本的判别正确率在大坝上、下游断面均为100%,非甲藻水华样本的判别正准确率在大坝上、下游分别为75.86%和82.00%(表5).使用本月理化数据和下月倪氏拟多甲藻密度建立的模型,甲藻水华样本的判别正确率在大坝上、下游断面分别为43.75%和11.11%,非甲藻水华样本的判别正确率在大坝上、下游分别为97.67%和100%.

从评价结果来看,模型3、4相比模型1、2对于甲藻水华样本的判别正确率大幅下降.一方面,适宜的水文条件、气象条件和营养条件是水华暴发的必要因子[5],通过Spearman相关性分析可以看出,相对于模型1、2,模型3、4中水温与倪氏似多甲藻密度的相关系数明显增大,由于嘉陵江甲藻水华多发生于春季的3-4月,而重庆地区春季温度变化较大,因此较长的预测周期增加了其不确定性,从而导致预测结果正确率的明显下降.另一方面,多数淡水甲藻可在不利条件下形成孢囊,因此根据孢囊的形成和释放周期选择适当的预测周期也是提高模型准确率的必要条件[24],但由于目前缺乏详细的倪氏拟多甲藻生理学知识,无法选择恰当的预测周期也可能是导致模型3、4判别正确率下降的重要原因.模型3、4相比模型1、2对非甲藻水华样本的判别正确率虽然有所下降,但通过分析判断错误的样本,发现它们大多集中在警戒值附近,所以从实际应用角度来看模型1、2对于判断草街水库大坝上、下游当月甲藻水华具有较强的预警能力.

2.3 与BP人工神经网络方法比较

为验证支持向量机在机器学习方法中的优越性,本文利用BP人工神经网络法同样建立4个预测模型,预测分类结果见表5.通过比较两种不同方法的结果可以看出,使用BP人工神经网络法时,除了模型2,其余3个模型的总样本判别正确率相比支持向量机法均明显下降,并且模型2和模型4的甲藻水华样本判别正确率均为0%,而非甲藻水华样本判别正确率为100%,说明模型训练后不具有很好的泛化能力,最终将检验组样本全部分为一类.由于模型2和模型4的训练样本数仅有143组和135组,并且甲藻水华样本数占总训练样本数的比例均不到5%,所以较少的训练样本以及某一类样本所占比例较低可能是导致这一结果的主要原因,熊秋芬等在进行支持向量机与神经网络方法比较研究中同样也发现,在样本数较少,并且某一样本数所占比例较低的情况下,SVM方法优势更明显[25].

表5 支持向量机分类模型预测结果

Tab.5 The classification results of prediction models

预测项目	支持向量机分类模型				BP人工神经网络分类模型			
	模型1	模型2	模型3	模型4	模型1	模型2	模型3	模型4
总样本判别正确率/%	80.00	84.75	89.22	87.30	71.43	84.75	83.33	85.71
甲藻水华样本判别正确率/%	100	100	43.75	11.11	83.33	0	37.50	0
非甲藻水华样本判别正确率/%	75.86	82.00	97.67	100	68.97	100	91.86	100

3 结论

1)通过相关性分析和PCA分析,筛选出影响甲藻水华最主要的环境变量建立甲藻水华预警模型,不仅大大缩短了构建模型所需的计算时间,还减少了水华预警所需要的理化因子,节约了环境管理部门监测成本.

2)使用当月理化数据预测下月甲藻水华正确率过低,所以进一步深入研究倪氏拟多甲藻的生理过程以及孢囊的形成萌发周期,可为选择甲藻水华的预测周期提供科学依据.

3)经实例应用表明运用支持向量机分类模型进行甲藻水华预警是可行的,该方法仅需将相应的理化数据提供给软件,利用计算机分析计算就可获得预测结果,而在BP人工神经网络模型中,除了网络结构、各层次节点、初始权重等选择很大方面要依靠研究者的个人经验,人为干预较多,并且训练模型对数据的要求更加严格,所以与其相比SVM方法更加简便、快捷,适于推广.

4 参考文献

[1] 徐恒省,洪维民,王亚超等.太湖蓝藻水华预警监测技术体系的探讨.中国环境监测,2008,24(2):62-65.

[2] 王长友,于 洋,孙运坤等. 基于 ELCOM-CAEDYM 模型的太湖蓝藻水华早期预测探讨. 中国环境科学,2013,**33**(3):491-502.
[3] 朱 利,王 桥,吴传庆等. 巢湖水华遥感监测与年度统计分析研究. 中国环境监测,2013,**29**(2):162-166.
[4] 周火艳,王崇云,彭明春等. 滇池水华分形结构动态研究. 环境科学与技术,2011,**34**(2):32-35.
[5] 朱爱民,乔 晔,梁友光等. 三峡水库支流童庄河拟多甲藻水华的监测. 水生态学杂志,2012,**33**(4):49-53.
[6] 姚绪姣,刘德富,杨正健等. 三峡水库香溪河库湾冬季甲藻水华生消机理初探. 环境科学研究,2012,**25**(6):645-651.
[7] 孔繁翔,马荣华,高俊峰等. 太湖蓝藻水华的预防、预测和预警的理论与实践. 湖泊科学,2009,**21**(3):314-328.
[8] 周立国,冯学智,王春红等. 太湖蓝藻水华的 MODIS 卫星监测. 湖泊科学,2008,**20**(2):203-207.
[9] 尚琳琳,马荣华,段洪涛等. 利用 MODIS 影像提取太湖蓝藻水华的尺度差异性分析. 湖泊科学,2011,**23**(6):847-854.
[10] Cha Y, Park SS, Kim K *et al*. Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. *Water Resources Research*, 2014,**50**(3):2518-2532.
[11] Cho S, Lim B, Jung J *et al*. Factors affecting algal blooms in a man-made lake and prediction using an artificial neural network. *Measurement*, 2014,**53**:224-233.
[12] 杨 敏,毕永红,胡建林等. 三峡水库香溪河库湾春季水华期间浮游植物昼夜垂直分布与迁移. 湖泊科学,2011,**23**(3):375-382.
[13] 汤宏波,胡 圣,胡征宇等. 武汉东湖甲藻水华与环境因子的关系. 湖泊科学,2007,**19**(6):632-636.
[14] 杨正健,刘德富,易仲强等. 三峡水库香溪河库湾拟多甲藻的昼夜垂直迁移特性. 环境科学研究,2010,**23**(1):26-32.
[15] 边归国. 九龙江北溪拟多甲藻水华防治与应急处置. 中国环境管理,2012,(1):45-49.
[16] 龙胜兴,陈 椽,俞振兴等. 贵州黔东南州三板溪水库春季拟多甲藻水华特征. 中国环境监测,2012,**28**(6):27-31.
[17] Vapnik V. The nature of statistical learning theory. New York: Springer, 1995.
[18] Xie Z, Lou I, Ung WK *et al*. Freshwater algal bloom prediction by support vector machine in Macau Storage Reservoirs. *Mathematical Problems in Engineering*, 2012, **2012**:1-12.
[19] Gokaraju B, Durbha SS, King RL *et al*. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2011,**4**(3):710-720.
[20] 国家环境保护总局《水和废水监测分析方法》编委会. 水和废水监测分析方法:第 4 版. 北京:中国环境科学出版社, 2002.
[21] 李正最,谢悦波. 洞庭湖富营养化支持向量机评价模型研究. 人民长江,2010,**41**(10):75-78.
[22] 周 鹏,曾 晖,周 原等. 支持向量机用于芳烃类化合物对芳烃受体亲和性 QSAR 研究. 环境科学学报,2006,**26**(1):124-129.
[23] 高 隽. 人工神经网络原理及仿真实例. 北京:机械工业出版社,2003.
[24] 张 琪,缪荣丽,刘国祥等. 淡水甲藻水华研究综述. 水生生物学报,2012,**36**(2):352-360.
[25] 熊秋芬,胡江林,陈永义. 天空云量预报及支持向量机和神经网络方法比较研究. 热带气象学报,2007,**23**(3):255-260.