# 基于随机森林分类算法的巢湖水质评价

张 颖 高倩倩
(上海海事大学信息工程学院,上海 201306)

**摘 要** 基于监测数据及机器学习算法的湖泊水质实时评价技术对当前湖泊水资源的管理、维护和保护具有重要意义。本文针对巢湖水质的类别评价,利用随机森林(Random Forest, RF)分类算法对该区域水质进行类别判定。与其他算法相比,随机森林算法有着精度高、可容忍噪声强等诸多优点。测试结果表明,当决策树的棵数 ntree = 300,分裂属性集中属性个数 mtry = 2 时,在合肥湖滨监测断面水质分类准确率可达 96.15%,在巢湖裕溪口监测断面水质分类准确率高达 100%,该方法具有稳健性较高、实用性强、泛化性能好等特点,能够有效进行水质评价。

**关键词** 随机森林算法 决策树 分裂属性集 水质评价

**中图分类号** TV213.4 **文献标识码** A **文章编号** 1673-9108(2016)02-0992-07

# Water quality evaluation of Chaohu Lake based on random forest method

Zhang Ying Gao Qianqian
(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

**Abstract** Real time evaluation of water quality based on monitoring data and machine learning algorithm has great significance for management, maintenances and protection of water resources in lake. Aiming at the class evaluation of water quality of Chaohu, a classification algorithm named random forest was used to determine the category of the water quality of this area. Comparing with other typical machine learning methods, this method has higher precision of classification and better tolerableness of noise. The testing result shows that when the quantities of the decision-making tree: ntree = 300 and the number of attributes of split attribute sets: mtry = 2, the accuracy rate of water quality classification in Hefei Hubin monitoring section could reach 96.15%, and it reaches as high as 100% in Yu Xikou monitoring section. The suggested method has higher robustness, stronger practicability and higher generalization performance. It can effectively fulfill water quality assessment with high precision.

**Key words** random forest algorithm; decision-making tree; split attribute sets; water quality assessment

随着国家进一步贯彻和落实经济社会的可持续发展战略,保护和利用水资源工作日益备受关注,国家也把水资源保护工作提高到了极为重要的位置。以湖泊为例,作为重要的国土资源,湖泊是自然生态系统的重要组成部分,是大自然赐予人类的“天然宝库”,在经济社会可持续发展中发挥着重要的作用。过去30年,随着湖泊流域人口压力的不断增大,工业化和现代化进程的加快,湖泊的开发利用与社会经济发展之间的矛盾日益加剧,已然成为制约可持续发展的瓶颈。水质状况恶化、富营养化加重等一系列的水质问题日益显露[1]。湖泊与人类的生存发展乃至区域生态平衡息息相关,湖泊的水质状况直接影响到沿湖居民的饮用水质量及该地区社会和经济的发展。因此,对于水质优劣情况的监测评价显得尤为重要,对水质的分析是对水环境的定性研究,即对水体进行有效的类别判定。

水质评价的研究自开展以来备受国内外学者的广泛关注。传统的统计学习要求样本数据为无穷大,然而一般的分类问题样本数目却往往有限。传统的综合评价方法有特征法、营养状态指数法、物元分析法和生物指标法等,但这些方法大多适用于水

基金项目:国家自然科学基金资助项目(61273068);上海市自然科学基金资助项目(12ZR1412600);上海市教委科研创新资助项目(13YZ084)
收稿日期:2014-10-17;修订日期:2014-11-26
作者简介:张颖(1968—),男,博士,副教授,主要从事智能信息处理、海洋环境多传感器信息融合和传感器网络等研究工作。E-mail: yingzhang@shmtu.edu.cn.

质结构和水生生态环境简单的水域,对于较为复杂的环境评价其效果较差。近年来,部分学者对水质提出一些新的评价模型,比如单因子评价法、模糊评价法、投影寻踪算法、人工神经网络和支持向量机等方法。这些方法基本上可以反映出水质情况,但是由于水质污染的随机性、污染物的复杂性以及评价因子和水质间的强非线性关系等原因,它们也都暴露出一些不尽如人意的缺点,尤其是模型的结构较复杂,调节参数较多,通用性较差等。

随机森林是文献[2]提出的一种集成学习算法,其实质是由许多随机生成的决策树组成,因此也叫做随机决策树。随机森林分类算法具有很好的分类性能,它能由有限的训练集样本得到较小的误差及较高的分类准确率,同时计算简单、建模方便、训练时间短、通用性强。随机森林算法提出之后,由于其良好的性能表现,该算法被广泛应用到诸多领域。在生物信息领域,Chen 等[3]利用 RF 算法进行了蛋白质相互作用的研究; Smith 等[4]利用判别分析法与 RF 算法对细菌源追踪数据进行对比研究。在经济管理领域,Ying 等[5]以银行客户的数据为例运用 RF 算法对客户流失情况进行了研究。在医学研究方面,Lee 等[6]利用 RF 算法通过利用肺部 CT 图像对肺结核进行检测; Ward[7]等利用 RF 算法的分类结果对红斑狼疮患者的死亡率进行短期预测取得较好结果。在其他方面,RF 算法也展现出了其优越性。孟杰[8]建立了基于随机森林分类算法的我国上市公司财务失败预警模型,通过与逻辑回归模型、支持向量机回归模型、CART 分类树模型和神经网络模型的预测结果的对比表明,随机森林模型的预测精度更高,稳健性更好,可以更好地提高我国上市公司的抗风险能力。康有等[9]以汉中盆地平坝区为例将 RF 算法应用于区域水资源可持续利用评价中,相比于支持向量机等算法,该方法实用性强、稳健性较高、泛化性能好。张雷等[10]以云南松分布模拟为例探究了随机森林算法在生态学中的应用。然而在水质评价方面,随机森林算法的研究和应用还不够充分。本文尝试将随机森林分类算法应用到巢湖水质类别判定中以提高评价效率和准确率。

## 1　研究区概况

巢湖,俗称焦湖,相传因有巢氏居住得名,亦说因湖呈鸟巢状而得名。巢湖处于皖中江淮之间,坐落于长江中下游北岸,属长江下游左岸重要水系,是安徽省境内最大湖泊。巢湖属于外流区域的湖泊,与外流河流相通,湖水能流进也能排出,含盐分少,为淡水湖,是中国第五大淡水湖,同时也是沿湖主要城镇 300 多万居民饮用水源之一。巢湖形成较为久远,经过漫长岁月演变,湖面逐渐萎缩至目前的 780 $km^2$。由安徽省巢湖管理局发布的巢湖概况显示,巢湖东西长 54.5 km,南北宽 21 km,巢湖流域面积 13 486 $km^2$,人口 1 020 万,流域涵盖 5 市 14 个县市(区)。巢湖水系发达,交通便捷,入湖河流 35 条,接纳南、北、西三面来水,呈向心状分布。注入巢湖水量最大的河流是丰乐河,占总径流量的 61.5%,裕溪河是惟一的出水通道,通往长江。

2014 年 7 月 7 日,安徽省环保局发布《2013 年安徽省环境状况公报》显示,巢湖水体平均水质为Ⅳ类、轻度污染。其中,东半湖水质为Ⅳ类、轻度污染; 西半湖水质为Ⅴ类、中度污染。由此知巢湖水质监控及评价工作刻不容缓,采取有效措施对巢湖水质进行准确评价,可以及时提醒巢湖流域相关部门对其采取有力措施,避免水质的进一步恶化。

## 2　数据来源及随机森林算法

### 2.1　数据来源

我国在全国范围内的主要水系都建有水质自动监测站,截至目前共建有 145 个重点断面水质自动监测站,可检测到的指标共有 8 项。本文选用巢湖两大主要监测断面即合肥湖滨监测断面和巢湖裕溪口监测断面。两监测断面分布图信息及巢湖主要水系如表 1 和图 1 所示。

表 1　巢湖流域监测断面信息

Table 1　Information of Chaohu monitoring sections

| 水 系 | 点位名称 | 河流名称 | 断面状况 |
|---|---|---|---|
| 巢 湖 | 安徽合肥湖滨 | 淮 河 | 湖体(西半湖) |
| | 安徽巢湖裕溪口 | 长 江 | 湖体(东半湖) |

本研究样本数据来自中华人民共和国环保部数据中心发布的《全国主要流域重点断面水质自动监测周报》。选取酸碱度(pH)、溶解氧(DO)、氨氮($NH_3$-N)、5 日生化高锰酸盐指数($COD_{Mn}$)为分析指标,截取 2013 年上半年前 26 周水质数据信息如表 2 所示。根据《地表水环境质量标准》(GB 3838-2002)对研究区内各主要断面的水体质量展开评价研究,这些参数都可以表征水体污染程度,依据这些参数对水质类别进行判定是准确和科学的[11]。采


**图 1 研究区监测断面分布图**

Fig. 1 Distribution of monitoring sections in the study area

集从 2010—2013 年 206 周的巢湖水质数据作训练样本数据，以 2014 年上半年的前 26 周作测试样本数据集对巢湖两大断面进行水质类别判定。

**表 2 2013 年上半年前 26 周水质数据信息**

**Table 2 Information of water quality in the first 26 weeks of 2013**

| 周 数 | pH | DO(mg/L) | $COD_{Mn}$(mg/L) | $NH_3$-N(mg/L) |
|---|---|---|---|---|
| 1 | 7.16 | 12.7 | 4.6 | 0.22 |
| 2 | 6.88 | 12.9 | 5.2 | 0.08 |
| 3 | 6.68 | 11.1 | 5.5 | 0.08 |
| 4 | 6.77 | 11.1 | 6.5 | 0.1 |
| 5 | 6.81 | 9.79 | 6.6 | 0.53 |
| 6 | 6.65 | 9.09 | 6.3 | 0.42 |
| 7 | 6.99 | 13 | 5.3 | 0.21 |
| 8 | 6.86 | 12.1 | 4.9 | 0.18 |
| 9 | 6.7 | 10.4 | 3.4 | 0.26 |
| 10 | 6.88 | 9.93 | 3.6 | 0.18 |
| 11 | 7.09 | 11 | 3.8 | 0.16 |
| 12 | 6.7 | 9.2 | 4 | 0.39 |
| 13 | 6.7 | 8.7 | 4.9 | 0.28 |
| 14 | 6.68 | 7.72 | 4.9 | 0.16 |
| 15 | 7.18 | 7.88 | 3 | 0.22 |
| 16 | 7.29 | 7.78 | 5.6 | 0.25 |
| 17 | 7.42 | 7.96 | 6.2 | 0.39 |
| 18 | 7.37 | 6.45 | 3.4 | 0.14 |
| 19 | 7.38 | 6.06 | 2.3 | 0.43 |
| 20 | 7.34 | 6.56 | 3.7 | 0.22 |
| 21 | 7.57 | 6.86 | 3.6 | 0.31 |
| 22 | 6.98 | 2.34 | 5.9 | 2.04 |
| 23 | 7.11 | 4.04 | 6.3 | 1.32 |
| 24 | 7.11 | 4.54 | 5.6 | 0.43 |
| 25 | 7.79 | 6.96 | 5.6 | 0.23 |
| 26 | 7.19 | 5.53 | 4.3 | 0.18 |

## 2.2 随机森林算法

随机森林算法是由美国加州大学伯克利分校统计系教授 Leo Breiman 等于 2001 年提出的一种基于 CART(classification and regression tree) 决策树的组合分类模型—随机森林，该算法是一种有监督的机器学习算法，同时也是一种现代分类和回归技术。该算法主要融合了两大随机化思想即 Ho 在 1998 年提出的随机子空间(random subspace) 思想[12] 和 Breimans 在 1996 年提出的 Bagging 思想，Bagging 思想又称为自助聚集(Bootstrap aggregating)[13]。Ho 所提出的随机子空间思想是在对决策树的每个节点进行分裂时，从全部特征变量属性集中随机等概率地抽取一个属性集，再从这个子集中选择一个最优属性来分裂节点。Bagging 算法是最早的集成学习算法，该算法与随机森林算法均是基于 Bootstrap 方法重采样产生多个训练集。Bootstrap 法重采样是指将数据样本集进行 N 次有放回的随机采样，抽取与原始样本数目一样的数据，得到 bootstrap 样本集。Bootstrap 法重采样示意图如图 2 所示。

随机森林算法的基本分类单元是决策树，该算法实质是一个包含多个决策树的分类器，并且其输出类别由决策树输出类别的众数而定。该算法要求调解参量少，运算效率高且能够处理高维(特征变量多) 数据，训练速度快而不会出现过拟合现象。随机森林算法对特征选取具有较好的鲁棒性，无需特征筛选也能得到较高的准确率，因此适用于超高维特征向量空间，同时随机森林算法对异常值和噪声具有较高的容忍度且具有较好的数据推广和范化能力[14]。

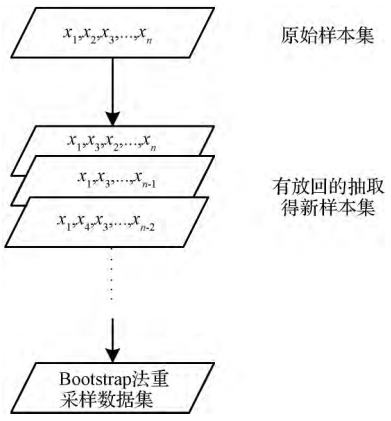


**图 2 Bootstrap 法重采样示意图**

Fig. 2 Diagram of Bootstrap resample

### 2.2.1 随机森林的构建

随机森林是以 $K$ 个决策树 $\{h(X,\Theta_k),k=1,2,\ldots,K\}$ 为基本分类单元，进行集成学习后得到的一个组合分类器。参数集 $\{\Theta_k,k=1,2,\ldots,K\}$ 是一个独立同分布的随机向量，它是由随机森林的两大随机化思想决定的。随机森林通过自助聚集法随

机选择样本生成决策树，每一棵决策树之间是没有关联的，而且每棵树都会完整生长而不会进行剪枝，并且在生成树的时候，每个节点的属性变量值都仅仅在随机选出的少数几个属性子集中产生。通过这 2 种在数据和属性变量值中的随机性可生成大量的树，我们称之为“随机森林”，如图 3 和图 4 所示。由于构建每个决策树时，随机抽取训练样本集和属性子集的过程是独立的，且总体是一样的，

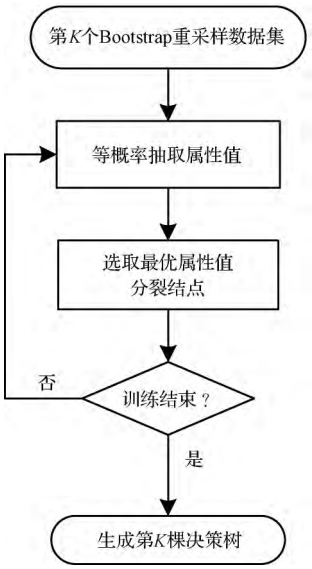


图 3　决策树的生成

Fig. 3　Generation of decision-making trees

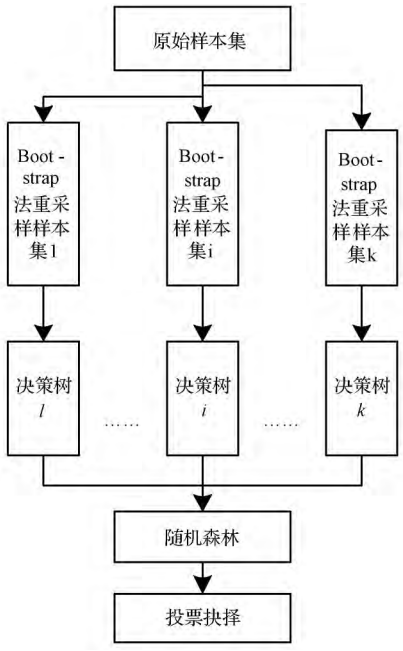


图 4　随机森林的生成

Fig. 4　Generation of random forest

因此参数集 $\{\Theta_k, k = 1, 2, \ldots, K\}$ 是一个独立同分布的随机向量[15]。

在得到森林之后，当有一个新的测试样本进入随机森林时，其实就是让每一棵决策树分别进行投票抉择，最终取所有决策树中输出类别最多的那类为分类结果。最终分类决策可用如下公式表示:

$$H(x) = \arg\max_{Y} \sum_{i}^{k} I(h_i(x) = Y)$$

式中: $H(x)$ 表示分类组合模型，$h_i$ 是单个决策树分类模型，$I(\cdot)$ 为示性函( 所谓示性函数是指一个函数使得当集合内有此数时值为 1，当集合内无此数时值为 0)，$Y$ 表示目标变量( 或称输出变量)[16]。

### 2.2.2　随机森林分类算法的模型建立

随机森林算法最重要的两个参数是决策树的棵数 ntree 和分裂属性集中属性个数 mtry，其中 mtry 个不同的输入属性在随机森林算法中是随机地选取某一属性值，且算法仅在此范围内选取最有效的结点分裂属性。分裂属性集中属性个数 mtry 是对随机森林分类效果影响较为敏感的参数，它的值可以自行设置。随机森林算法流程可简述如下:

(1) 利用 Bootstrap 法重采样原始数据样本集 $X$，随机生成 K 个训练样本集 $X_1^*, X_2^*, \ldots, X_K^*$;

(2) 利用每个生成的训练集，生成对应的决策树 $T_1, T_2, \ldots, T_k$，在每个中间节点( 非叶子节点) 上选择 mtry 个属性( 从 $M$ 个属性集中随机抽取的分裂属性集) 中最佳分裂方式的属性作为当前节点的分裂属性在此节点上进行分裂;

(3) 每棵决策树都完整生长而不进行剪枝;

(4) 将每棵决策树对原始数据样本集 $X$ 进行测试分类;

(5) 采用投票的方式，将 $K$ 棵决策树输出最多的类别作为原始数据样本集 $X$ 的所属类别。

森林中决策树的构建是模型建立的核心，每棵决策树都最大限度地生长，没有剪枝。随机森林分类算法中不同的决策树个数对模型泛化性能也存在一定的影响。决策树个数的多少直接影响随机森林分类算法的运算速度和分类效果，因此决策树的个数对建模至关重要。若决策树的棵数太多，则会导致随机森林算法的速度下降; 若决策树的棵数太少，则会导致模型的分类准确率下降。因此，我们需要采取折中的方法合理地确定随机森林分类模型决策树的棵数。但是，随机森林分类算法的优势在于它的训练速度快，算法简洁，效果稳定，网络只需要训

练一次便可获得理想结果。因此，不能够再将确定决策树棵数的迭代运算集成到训练模型内，而是应该提前确定和设置合理的决策树棵数。

由于具有相同复杂度和样本个数的数据所需设置的决策树的棵数基本相同。所以，可以通过 400 组水质数据进行样本训练来提前确定决策树的棵数。为了减少决策树棵数随机性的影响，可以作如下处理：确定决策树的棵数以后，建立 100 个随机森林分类模型，然后，取水质分类准确率的平均值作为当前决策树棵数下随机森林分类的准确率。某次运行结果如图 5 所示。

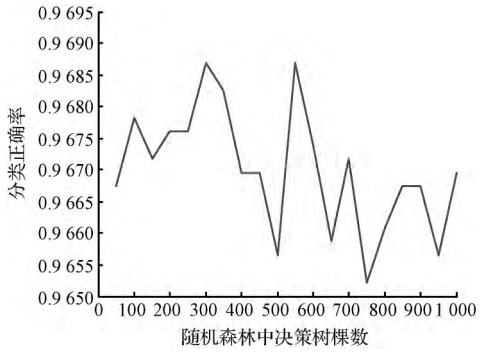


图 5 随机森林中决策树棵数对分类正确率的影响

Fig. 5 Impact of number of decision-making trees in random forest on accuracy

从图 5 可以看出，针对巢湖水质的数据集而言，随着决策树棵数的增加，随机森林的分类正确率并未持续升高，相反出现了一些曲折，综合考虑随机森林中包含的决策树的棵树与建模的速度，选取随机森林分类算法中包含 300 棵决策树是比较理想的。

## 3 结果与分析

根据上述获取的样本数据，用随机森林分类算法进行水质评价模型建模，其中，决策树的棵数为 300。分裂属性集中属性个数的值可以自行设置，以合肥湖滨监测断面水质为例，mtry = 1 时，分类准确率为 24/26; mtry = 2，3 时分类准确率为 25/26; mtry = 4 时分类准确率为 24/26; 此外结合一般情况下 mtry = $\lfloor\sqrt{M}\rfloor$（其中 M 为总的属性个数，符号 $\lfloor\bullet\rfloor$ 表示向下取整），故此时分裂属性集中属性个数 mtry = 2。使用 2010 年至 2013 年的 206 周水质样本作训练组数据对模型进行训练，然后，用训练好的随机森林评价模型对巢湖两监测断面的 2014 年前半年 26 周水质进行类别判定，其结果如图 6 所示。

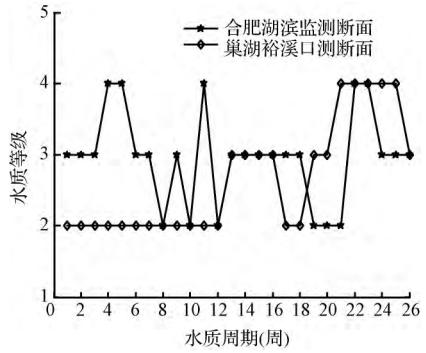


图 6 随机森林分类算法水质类别判定结果

Fig. 6 Results of water quality classification based on random forest classification algorithm

从水质类别判定结果看，巢湖裕溪口监测断面的水质在前 12 周水质都达到 II 级水质的指标，其水质情况均优于合肥湖滨监测断面水质，在接下来的 14 周内巢湖裕溪口监测断面的水质与合肥湖滨监测断面水质相差不大，但也略优于后者。而在实际情况中，由于巢湖裕溪口的裕溪河是惟一的通往长江的出水通道，与外界流通性好，水质也相对于合肥湖滨监测断面略优，故其评价结果与实际情况相符。巢湖两监测断面真实水质等级如图 7 所示。

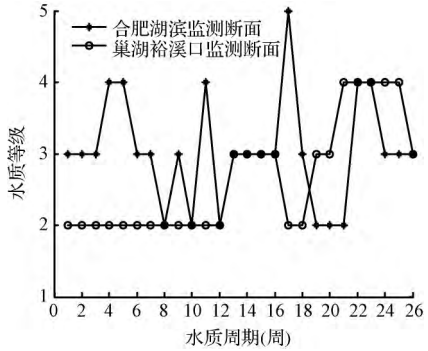


图 7 实际水质等级

Fig. 7 Actual water quality grade

对比图 6 和图 7 所示可知，运用随机森林构建的水质类别判定模型，在对巢湖水质评价中取得了优良的效果，模型评价水质等级变化趋势同真实指标数据变化趋势相符，说明了模型评价的有效性。

为了直观地分析该方法在水质分类中的效果，将极限学习机（extreme learning machine，ELM）算法和支持向量机（support vector machine，

SVM) 算法分别建模进行比较。针对相同的输入样本数据,应用 ELM 和 GA-SVM 进行软测量建模的分类结果对比及运行时间对比如图 8、图 9 和表 3 所示。

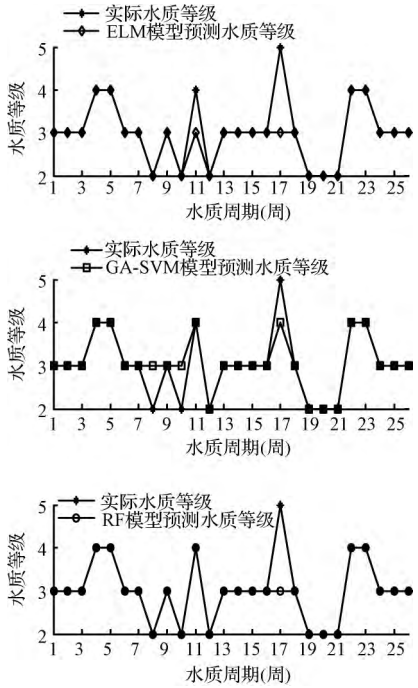


**图 8　安徽合肥湖滨监测断面水质类别判定对比图**

Fig. 8　Classification of water quality in Hefei Hubin monitoring section of Anhui Province

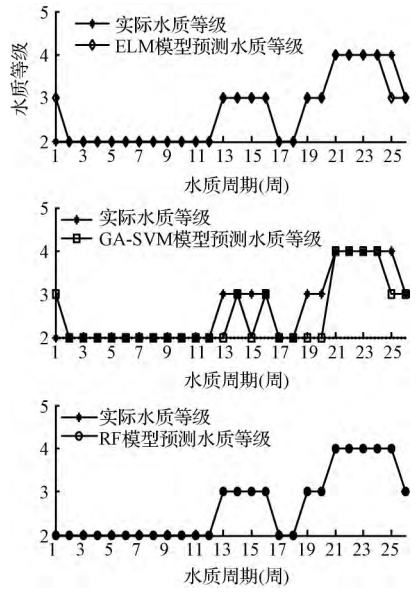


**图 9　巢湖裕溪口监测断面水质类别判定对比图**

Fig. 9　Classification of water quality in Chaohu Yu Xikou monitoring section

**表 3　分类准确率及运行时间对比**

**Table 3　Comparison of classification accuracy and running time**

| 分类准确率和时间监测断面 | RF | ELM | GA-SVM |
|---|---|---|---|
| 合肥湖滨监测断面 | 25/26 (96.15%) | 24/26 (92.31%) | 23/26 (88.46%) |
| | 0.290 069 s | 0.067 776 s | 115.486 906 s |
| 巢湖裕溪口监测断面 | 26/26 (100%) | 24/26 (92.31%) | 20/26 (76.92%) |
| | 0.298 428 s | 0.067 285 s | 68.564 053 s |

综合图 8、图 9 和表 3 对比可以看出: RF 分类模型在合肥湖滨监测断面除第 17 周水质评价错误外,其余各周水质评价完全正确,正确率达96.15%; 在巢湖裕溪口监测断面,26 周水质评价完全与实际相符,准确率高达 100%; RF 和 ELM 分类模型运行所需时间都不足一秒,而 GA-SVM 分类模型所需时间几十秒甚至百秒以上; 虽然 ELM 分类模型的运行时间比 RF 分类模型更短,但在分类准确率上却略微逊色; GA-SVM 分类模型无论在准确率还是运行时间上都不如 RF 分类模型出色。RF 分类模型的分类准确率较高,在水质评价中体现了较强的分类能力和抗干扰能力,具有较好的泛化性能,稳定性较好; ELM 分类模型的分类准确率次之,而 GA-SVM 分类模型较差。此评价结果显示,基于随机森林算法的水质类别判定模型适合巢湖水质类别的判定,可以对巢湖水质状况做出合理有效的评价。

## 4　结　论

本文通过使用公开可查的巢湖水质监测数据,利用随机森林模型的特征及其在分类中的适用性,建立了基于随机森林模型的巢湖水质类别判定模型。经过决策树的选择,表明所选决策树的棵数能够较好地预测水质等级,即该模型能有效地对巢湖整体水质作出合理的评价。此外,文中还利用 ELM 和 GA-SVM 算法构建了评价模型,3 种模型的预测分类结果表明,使用随机森林模型得到的水质类别判定结果要优于 GA-SVM 和 ELM 两种算法。随机森林模型解决了其他算法稳健性不足和过学习等问题,对数据前提条件的要求宽松,算法简洁,调节参数少,训练速度快,因此其在水质评价分析中值得推广,应用前景将非常广阔。

## 参 考 文 献

[1] 刘鸿亮. 湖泊富营养化控制. 北京: 中国环境科学出版

社，2011

[2] Breiman L. Random forests. Machine Learning，**2001**，45(1)：5-32

[3] Chen Xuewen，Liu Mei. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics，**2005**，21(24)：4394-4400

[4] Smith A.，Sterba-Boatwright B.，Mott J. Novel application of a statistical technique，random forests，in a bacterial source tracking study. Water Research，**2010**，44(14)：4067-4076

[5] Ying Weiyun，Li Xiu，Xie Yaya，et al. Preventing customer churn by using random forests modeling//Proceedings of the IEEE International Conference on Information Reuse and Integration(IRI 2008). Las Vegas，NV，USA：IEEE，**2008**：429-434

[6] Lee S. L. A.，Kouzania A. Z.，Hu E. J. Random forest based lung nodule classification aided by clustering. Computerized Medical Imaging and Graphics，**2010**，34(7)：535-542

[7] Ward M. M.，Pajevic S.，Dreyfuss J.，et al. Short - term prediction of mortality in patients with systemic lupus erythematosus：Classification of outcomes using random forests. Arthritis Care & Research，**2006**，55(1)：74-80

[8] **孟杰. 随机森林模型在财务失败预警中的应用. 统计与决策，2014，**(4)：179-181

[9] **康有，陈元芳，顾圣华，等. 基于随机森林的区域水资源可持续利用评价. 水电能源科学，2014，**32(3)：34-38
Kang you，Chen Yuanfang，Gu Shenghua，et al. Assessment of sustainable utilization of regional water resources based on random forest. International Journal Hydroelectric Energy，**2014**，32(3)：34-38(in Chinese)

[10] **张雷，王琳琳，张旭东，等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例. 生态学报，2014，**34(3)：650-659
Zhang Lei，Wang Linlin，Zhang Xudong，et al. The basic principle of random forest and its applications in ecology：A case study of *Pinus yunnanensis*. Acta Ecologica Sinica，**2014**，34(3)：650-659(in Chinese)

[11] **席北斗，赫英臣，龚斌. 德国巴伐利亚州水域水质分类特征. 人民黄河，2010，**32(1)：50-51

[12] Ho T. K. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence，**1998**，20(8)：832-844

[13] Breiman L. Bagging predictors. Machine Learning，**1996**，24(2)：123-140

[14] **马昕，王雪，杨洋. 基于随机森林算法的大学生异动情况的预测. 江苏科技大学学报(自然科学版)，2012，**26(1)：86-90
Ma Xin，Wang Xue，Yang Yang. Prediction of degradation for undergraduate using random forest. Journal of Jiangsu University of Science and Technology (Natural Science Edition)，**2012**，26(1)：86-90(in Chinese)

[15] **董师师，黄哲学. 随机森林理论浅析. 集成技术，2013，**2(1)：1-7
Dong Shishi，Huang Zhexue. A brief theoretical overview of random forests. Journal of Integration Technology，**2013**，2(1)：1-7(in Chinese)

[16] **方匡南，吴见彬，朱建平，等. 随机森林方法研究综述. 统计与信息论坛，2012，**26(3)：32-38
Fang Kuangnan，Wu Jianbin，Zhu Jianping，et al. A review of technologies on random forests. Statistics & Information Forum，**2011**，26(3)：32-38(in Chinese)