基于改进随机森林算法的通榆河水生态健康状况评价研究

刘 飞[1]，凌洪军[2]，路广宽[3]

（1. 江苏省建湖县沿河水利管理服务站，江苏 盐城 224700；2. 江苏省建湖县水利局，江苏 盐城 224700；
3. 江苏省建湖县通榆河管理所，江苏 盐城 224700）

摘要：通过改进传统的随机森林算法，对江苏省通榆河的水生态健康状况进行评价研究，建立了包含 18 个具体指标的通榆河水生态健康状况评价体系，并将评价结果分为“病态、微病态、亚健康、微健康、健康”5 个等级。模型改进结果表明，改进的随机森林算法（IRF）较传统的随机森林（RF）算法和径向基人工神经网络（ANN－RBF）在模型误差和计算效率上均有明显提升，算法性能得到较大改善；评价结果显示，通榆河在 2016～2018 年的水生态健康状况分别为微病态、亚健康、微健康，整体呈现良好的发展态势。改进的模型可为相关评价研究提供借鉴和指导。

关键词：改进随机森林算法；水生态健康；评价体系；通榆河

中图分类号：X826 **文献标识码**：B **文章编号**：1007-7839(2019)10-0018-05

Study on the evaluation of water ecological health status of Tongyu River based on improved random forest algorithm

LIU Fei[1], LING Hongjun[2], LU Guangkuan[3]

(1. *Yanhe Water Conservancy Management Service Station of Jianhu County*, *Yancheng* 224700, *Jiangsu*;
2. *Jianhu Water Conservancy Bureau*, *Yancheng* 224700, *Jiangsu*;
3. *Tongyu River Management Office of Jianhu County*, *Yancheng* 224700, *Jiangsu*)

Abstract: The water ecological health status of Tongyu River in Jiangsu Province was evaluated by improved random forest algorithm, the evaluation system of Tongyu River water ecological health status including 18 specific indicators was established, and the evaluation results were divided into 5 grades of morbidity, micro－morbidity, sub－health, micro－health, and health. The improved model results showed that IRF had obvious improvement in model error and computational efficiency compared with RF and ANN－RBF, and the algorithm performance was greatly improved. The evaluation results showed that the water ecological health status of Tongyu River in 2016－2018 was micro－pathological, sub－health and micro－health, and the overall development trend was good. The improved model could provide reference and guidance for relevant evaluation research.

Key words: improved random forest algorithm (IBF); water ecology health; evaluation system; Tongyu River

1 概述

通榆河是江苏省江水东引，江水北调工程项目的一部分，它南起南通市海安县、北至连云港市赣榆区，全长 375 km。是排涝、灌溉的重要通道，同时承担着贯穿南北航道的重要角色[1]。本研究的评

收稿日期：2019-06-17

作者简介：刘飞（1989—），男，本科，主要从事河道治理和评价研究等相关工作。

价范围为通榆河在盐城市建湖县境内的部分河段，评价河段南起潭洋河，北至八份河，建湖境内长度为 17.1 km，现状河底宽 40 ~ 70 m，底高程 -4 m，堤顶高程 4 ~ 5 m，通榆河标准 20 年一遇、排涝标准 10 年一遇。因通榆河对建湖县城市的发展起着重要的水资源支撑作用，其水生态健康状况势必会影响建湖县未来的发展[2]。因此，对近 3 年通榆河水生态健康状况进行评估，可以为进一步开展水资源管理和保护工作提供行动指南和理论参考。

2　随机森林算法及改进

2.1　随机森林算法

随机森林算法[3-5]（Random Forest，RF）是一种建立在统计学上，以组合分类为基础的智能算法，它具有较强的非线性模拟能力、泛化能力和数据挖掘能力。算法中需要人为设置的参数较少，如此可以减少评价研究问题的主观性。其基本原理是装袋算法和随机子空间算法的集合，基本单元为决策树，将多个决策树组合在一起形成森林，通过每个决策树分类预测投票，从而得出最终的分类及评价结果。随机森林的形成和评价示意如图 1 所示。

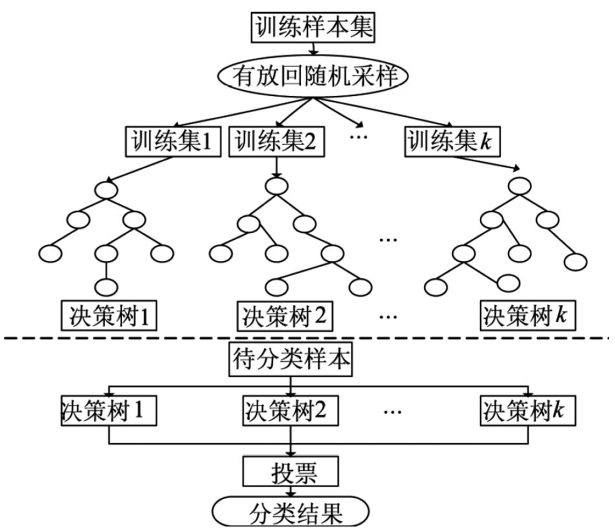

图 1　随机森林形成及样本评价示意图

对于每一棵决策树 $Ti(i=1,2,\cdots,nTree)$，有放回的随机抽样产生训练集 Di，若以 F 为每个节点划分的属性数量，为了搭建决策树 Ti，从每个节点当前可利用的属性集合中以随机选择的方式选取 F 个属性作为节点划分的待选属性集，然后以信息增益、信息增益率等指标为依据进行最佳分类属性的选择。

若当前样本集合 D 中第 i 类样本所占的比例为 P，则样本集合 D 的信息熵为：

$$Entropy(D) = -\sum_{i=1}^{c} P_i \log_2 P_i \tag{1}$$

在特征 A 作用后，样本集合 T 被分成 k 个部分。此时的信息熵、信息增益、信息值、信息增益率分别为：

$$Entropy(D_A) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} Entropy(D_A) \tag{2}$$

$$Gain(D,A) = Entropy(D) - Entropy(D_A) \tag{3}$$

$$SplitEntropy(D,A) = -\sum_{i=1}^{c} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|} \tag{4}$$

$$GainRation(D,A) = \frac{Gain(D,A)}{SplitEntropy(D,A)} \tag{5}$$

2.2　随机森林算法改进

随机森林算法在进行评价研究时，因待评价样本的特性，有时会存在数据比重严重不平衡的状况。此外，节点属性特性选择完全是随机化的，这样较容易导致比较重要的属性特征可能被过滤，最终造成算法模型欠拟合等问题。为次，本研究引入基于信息值的节点属性随机选择优化方案。

选择信息值计算属性特征的重要程度，首先计算每个非类别属性和类别属性之间的信息值，然后将值进行大小顺序的排列，相关性大的值也越大，对应的排列在序列的前面。其次，根据信息值将属性空间划分为两部分，分别为强相关和弱相关。在随机选择节点的属性时，在强和弱两个区间分别进行一定比例的选择。

其中，信息值（IV）是衡量属性 X 和目标类别属性 Y（Y 为二元属性）之间相关性的指标，计算式为：

$$IV = \sum_{i=1}^{n} (P_i - P_i')\ln \frac{P_i}{P_i'} \tag{6}$$

式中，n 为属性 X 的类别数量，P_i 为属性 $X=x_i$ 时目标类别 $Y=y$ 的概率，P_i' 为 $X=x_i$ 是 $Y\neq y$ 的概率。

3　河流水生态健康评价指标体系

3.1　评价指标筛选与确定

在进行河流水生态健康评价时，应当全面考虑影响水生态的各个层面和因素，进行归纳分类。指标的数量不宜过多，避免出现冗余重复信息，应当尽量体现不同维度。同时，要严格遵循科学性、独立性、代表性、规范性和可操作性原则。通过比较分析，本研究将影响河流水生态健康状况的因素分为 4 类，即水文特征、水质特征、水生态特征和社会服务特征。针对每个准则层的指标再进行指标筛选，最后通过筛选确定出了 18 个水生态健康评价指

标[6],对于每个健康指标的计算方法具体见表 1。

3.2 水生态健康分级标准

参考《江苏省健康河流诊断指数体系研究》和《盐城市生态水系规划报告》,对选取的 18 项评价指标的标准进行划分,共将健康状况分为健康、微健康、亚健康、微病态和病态 5 种等级,其对应的健康指数分别为 1,2,3,4,5,具体见表 2。

4 实例研究

将改进随机森林算法(IRF)应用于通榆河 2016~2018 年近 3 年的水生态健康评价研究,从而探究通榆河近 3 年的水生态健康状况,具体评价分为 5 个步骤:

a. 构造训练样本和检验样本。因为改进随机森林算法在进行样本生成时是采用的随机生成的方法,因此,首先需要在评价标准的各个阈值范围内,随机生成 300 组样本。每个健康状况内随机生成 60 组,5 个等级共计 300 组,其中选取 240 组作为训练样本,60 组作为检验样本,输出则是通过健康指数来控制,样本及输出模式见表 3。

b. 对评价指标数据进行预处理。为了避免因不同指标量纲和数量级给算法带来差异,需要对评

表 1 通榆河流水生态健康评价指标体系

目标层	准则层	指标层
河流水生态健康评价	水文特征	河岸植被覆盖度(C1)
		岸坡稳定性(C2)
		河道连通性(C3)
		断面平均流速(C4)
		生态流量保证率(C5)
	水质特征	水质综合指数(C6)
		底泥污染指数(C7)
	水生态特征	鱼类多样性(C8)
		浮游植物多样性(C9)
		底栖动物完整性(C10)
	社会服务功能	堤防功能达标率(C11)
		岸线利用率(C12)
		被利用岸线完好率(C13)
		公众满意度(C14)
		供水保证率(C15)
		水功能区水质达标率(C16)
		通航保证率(C17)
		水资源开发利用率(C18)

表 2　通榆河流水生态健康评价等级划分表

健康状况	健康	微健康	亚健康	微病态	病态
健康指数	1	2	3	4	5
C1	(90,100]	(80,90]	(65,80]	(50,65]	(0,50]
C2	(90,100]	(80,90]	(70,80]	(60,70]	(0,60]
C3	(0.9,1.0]	(0.75,0.9]	(0.6,0.75]	(0.45,0.6]	(0,0.45]
C4	≥1.0	(0.8,1.0]	(0.6,0.8]	(0.4,0.6]	(0,0.4]
C5	(95,100]	(90,95]	(80,90]	(75,80]	(0,75]
C6	1	2	3	4	5
C7	(0.9,1.0]	(0.75,0.9]	(0.6,0.75]	(0.45,0.6]	(0,0.45]
C8	(2.5,3.0]	(2.0,2.5]	(1.5,2.0]	(1.0,1.5]	(0,1.0]
C9	(2.5,3.0]	(2.0,2.5]	(1.0,2.0]	(0.5,1.0]	(0,0.5]
C10	(2.5,3.0]	(2.0,2.5]	(1.0,2.0]	(0.5,1.0]	(0,0.5]
C11	(97,100]	(90,97]	(80,90]	(65,80]	(0,65]
C12	(95,100]	(80,95]	(60,80]	(50,60]	(0,50]
C13	(90,100]	(80,90]	(70,80]	(50,70]	(0,50]
C14	(0.8,1.0]	(0.7,0.8]	(0.6,0.7]	(0.4,0.6]	(0,0.4]
C15	(97,100]	(90,97]	(85,90]	(70,85]	(0,70]
C16	(90,100]	(80,90]	(70,80]	(60,70]	(0,60]
C17	(97,100]	(90,97]	(80,90]	(70,80]	(0,70]
C18	(0,10]	(10,25]	(25,40]	(40,60]	(60,100]

表 3　河流健康评价样本及期望输出设计

样本组	目标输出健康指数	健康状况
1 ~ 60	1	健康
60 ~ 120	2	微健康
120 ~ 180	3	亚健康
180 ~ 240	4	微病态
240 ~ 300	5	病态

价指标进行归一化处理。由于有些指标是属于越大越好型，称为正向性指标，反之成为负向型指标。对于正向型指标和负向型指标，分别采用式(7)和式(8)进行归一化处理。本次研究中除了水质综合指数(C6)和水资源开发利用率(C18)外，其余均为正向性指标。

$$x' = (x - x_{min})/(x_{max} - x_{min}) \quad (7)$$

$$x' = (x_{max} - x)/(x_{max} - x_{min}) \quad (8)$$

式中,x 和 x' 为归一化前和后的指标值,x_{min} 和 x_{max} 分别人评价阈值的上限和下限。

c. 构建改进随机森林算法(IRF)模型。根据随机森林算法的原理,利用 MATLAB 软件和随机森林工具集合(Randomforest - matlab)构建 RF,并以基于信息值的节点属性随机选择对 RF 进行优化。以选取的经过归一化处理后的训练样本作为模型的输入数据,共计 18 项输入,以表 3 对应的样本组中的目标输出健康指数为模型输出数据,共计 1 项。由此构建 18 输入对应 1 输出的映射关系,而这个映射过程的处理就是改进随机森林算法的核心。为了检验改进之后的模型性能的好坏,以同样的方式建立人工径向基神经网络评价模型(ANN - RBF)和未改进的随机森林算法(RF)评价模型,对比分析 3 个模型的计算效果和性能。

d. IRF 模型的训练及精度检验。利用训练样本和检验样本对 ANN - RBF、RF 和 IRF 模型分别进行训练,以平均相对误差绝对值(AREA)、最大相对误差绝对值(MREA)和运行时间(RT)作为评价 3 个模型效果的性能指标,分别让 3 个模型行进 100 次,选取各个性能指标进行对比,其中运行时间为模型进行 100 次计算所消耗的 CPU 运算时间(每个模型采用相同的 CPU 数目)。为了达到最好的模型性能,需要进行各自模型参数的设置。影响随机森林的模型参数主要有:①决策树棵数数目;②分裂特征个数。而影响 ANN - RBF 模型的参数有:①径向基函数分布密度;②期望误差。为了使得各自的模型性能达到最优值,采用网格搜索法,经反复测试,对于 RF 和 IRF 模型,设置决策树棵数均为 1000,分裂特征集中的特征个数为输入特征的总维数的算数平方根。对于 ANN - RBF 模型,设置径向基函数分布密度和期望误差分别为 2 和 1/1000。通过 3 个模型的运行计算,将其性能指标进行对比,见表 4。

e. 模型性能评价分析。根据表 4 中 ANN - RBF、RF 和 IRF 模型的性能评价可知,经过改进之后的 IRF 模型在平均相对误差绝对值(AREA)、最大相对误差绝对值(MREA)和运行时间(RT)三个方面均明显优于 ANN - RBF、RF 模型,IRF 表现出较好的评价精度和泛化能力,且在运行速度上也有提升,可以进一步节省计算时间和计算机资源。因此,可以采用改进随机森林算法进行通榆河水生态健康评估。

f. 应用模型进行实例评价。通过收集整理得到通榆河 2016 ~ 2018 年评价指标数据,具体见表 5。对原始数据进行归一化处理,并以上述训练好的 IRF 模型进行评价分析,计算得到通榆河 2016 ~ 2018 年水生态健康状况。具体见表 6。

从通榆河 2016 ~ 2018 年水生态健康状况可以看出,通榆河在近 3 年整体表现出逐年变好的趋势,健康指数从 2016 年的 4.34(微病态)上升到了 2018 年的 2.19(微健康)。由此表明盐城市和建湖县政府及水利和环保等相关部门结合推行“河长制”、积极组织开展的“一河一策”治理等方面,对通榆河的上下游及左右岸河况进行了详尽排查,并出台了相应的治理方案并进行了监督实施,包括通榆河河岸植被绿化,修复老旧堤防岸线等,通过水生态健康状况研究显示,治理工作初显成效。

5 结论

随着“十九大”报告关于“坚持人与自然和谐共生,树立和践行绿水青山就是金山银山的理念,坚持节约资源和保护环境”基本国策的提出。河流水生态健康逐渐成为指导城市发展的重要因素。本文通过对现有的随机森林算法进行研究,提出了基于信息值的节点属性随机选择优化的改进随机森林算法(IRF),并采用该模型对江苏省盐城市建湖县境内的通榆河进行了水生态健康评价。通过评

表 4 ANN - RBF、RF 和 IRF 模型性能评价

模型	AREA/%		MREA/%		RT/s
	平均值	变化范围	平均值	变化范围	平均值	累积时间
ANN - RBF	3.47	2.81 ~ 4.13	17.55	12.76 ~ 33.34	1.54	74.31
RF	1.08	0.79 ~ 1.63	9.11	6.76 ~ 15.37	0.72	24.28
IRF	0.46	0.37 ~ 0.55	4.38	2.66 ~ 8.10	0.32	10.75

(下转第 68 页)

(上接第 22 页)

表 5　通榆河 2016～2018 年水生态评价指标

年份	C1	C2	C3	C4	C5	C6	C7	C8	C9
2016	55.20	64.70	0.37	0.15	77.10	3.00	0.67	1.62	1.42
2017	78.60	80.20	0.76	0.56	88.40	3.00	0.78	1.79	1.33
2018	81.40	85.60	0.76	1.50	89.00	3.00	0.76	2.18	1.98
年份	C10	C11	C12	C13	C14	C15	C16	C17	C1
2016	0.84	75.50	58.10	0.69	0.44	74.20	60.70	73.90	66.4
2017	1.55	82.60	76.30	0.85	0.91	90.40	95.40	96.80	42.3
2018	2.34	94.00	93.00	95.00	0.83	88.00	87.00	88.00	37.2

表 6　通榆河 2016～2018 年水生态健康状况

年份	健康指数	健康状况
2016	4.34	微病态
2017	3.27	亚健康
2018	2.19	微健康

价结果显示,通榆河在 2016～2018 年近 3 年的水生态表现出向好的趋势,水生态健康状况实现了微病态～亚健康～微健康的过程转变。研究表明,盐城市和建湖县在推进生态文明建设和“河长制”、“一河一策”等工作方面突显成效,良好的河流水生态健康环境可为今后城市的发展提供良好的水资源背景基础。

参考文献:

[1] 刘萍. 建湖县河湖管理范围划定工作研究[J]. 江苏水利, 2018(11):53－57.

[2] 镇云, 李友春. 城市河道生态护岸稳定分析及应用[J]. 江苏水利, 2018(06):66－69＋72.

[3] 鲁亚会. 基于随机森林特征选择的贝叶斯分类模型及应用[D]. 郑州:华北水利水电大学, 2017.

[4] 余胜男, 陈元芳, 顾圣华, 等. 随机森林在降水量长期预报中的应用[J]. 南水北调与水利科技, 2016, 14(01):78－83.

[5] 赖成光, 陈晓宏, 赵仕威, 等. 基于随机森林的洪灾风险评价模型及其应用[J]. 水利学报, 2015, 46(01):58－66.

[6] 单红喜. 基于模糊数学的河道健康综合评价方法研究[J]. 江苏水利, 2018(05):57－62.