随机森林与模糊识别耦合方法在河流健康评价中的应用

王亦斌[1]，孙 涛[1]，陈 谌[2]

（1. 南水北调东线江苏水源有限责任公司，江苏 南京 210029；
2. 南京工业大学 电气工程与控制科学学院，江苏 南京 211816）

摘要：随着现代社会对生态环境关注度的提高，而河流在生态环境中又起到了举足轻重的作用，如何准确构建河流健康评价模型成为了河流治理的关键内容。研究首先建立河流健康评价指标体系，利用随机森林 gini 指数计算出每个指标的权重，再构建指标相对隶属度函数，最后进行模糊识别综合计算，并将其应用到某市 6 条主要河流的健康评价中。试验结果表明：编号为 1～5 的河流评价结果均在[0.70，0.85]之间，评价等级为良；编号为 6 的河流评价结果在 0.70 以下，评价等级为差，说明本文方法的评价结果与实际河流评价情况基本一致。因此，本文提出的随机森林与模糊识别耦合方法合理、可行，不仅科学合理地确定了河流健康评价指标的权重，也提高了模型评价的准确性。

关键词：河流健康评价；相对隶属度；随机森林；模糊识别

中图分类号：X826 **文献标识码**：B **文章编号**：1007-7839(2019)05-0025-05

Application of random forest and fuzzy recognition coupling method in river health assessment

WANG Yibin[1], SUN Tao[1], CHEN Chen[2]

(1. *Jiangsu water source limited liability Company of South to North Water Transfer East line*, *Nanjing* 210029, *Jiangsu*; 2. *College of Electrical Engineering and Control Science*, *Nanjing Tech University*, *Nanjing* 211816, *Jiangsu*)

Abstract: With the increasing attention paid to the ecological environment in modern society, rivers play a decisive role in the ecological environment. How to accurately build river health evaluation model has become the key content of river governance. Firstly, the river health evaluation index system was established, the weight of each index by using gini index of stochastic forest was calculated. Then, the relative membership function of the index was constructed. Finally, the comprehensive calculation of fuzzy recognition was carried out and applies to the health evaluation of six main rivers in a city. The test results showed that the evaluation results of rivers numbered 1 – 5 were between [0.70, 0.85], and the evaluation grade was good; the evaluation results of rivers numbered 6 were below 0.70, and the evaluation grade was poor, which showed that the evaluation results of this method were basically consistent with the actual river evaluation situation. Therefore, the coupling method of random forest and fuzzy identification proposed was reasonable and feasible. It not only determined the weight of river health

收稿日期：2018-10-4

基金项目：国家自然科学基金青年基金项目(11801267)；国家重点研发计划项目(2017YFC1502603)

作者简介：王亦斌(1970—)，女，本科，高级工程师，主要从事水利工程建设与管理工作。

通讯作者：陈谌(1993—)，男，硕士，研究方向为智能算法、模式识别。

evaluation index scientifically and reasonably, but also improved the accuracy of model evaluation.
Key words: river health assessment; relative membership degree; random forest; fuzzy identification

1 概况

随着中国社会经济进入发展转型的关键阶段，保护生态环境、发展绿色经济成为了当今社会的共识。河流资源作为人类社会发展的基础，在调节气候、改善环境等方面都起到了重要的作用，近年来随着各种河流污染问题的加剧，越来越多的河流管理方法被提出，其中河流健康评价模型作为一种新颖的评价方法被广泛应用[1]。国内研究学者先后提出了许多河流健康评价方法，例如主成分分析法、模糊粗糙集法、突变级数法等。河流健康评价指标具有多层次多指标的特征，有一定的模糊性和复杂性，本文从三个层级构建指标体系，控制层包括社会和自然两个方面；准则层包括形态特征、环境特征、河道管理能力、供水能力等 8 个准则；指标层包括河岸稳定性、水质综合指数、防洪工程达标率、公众满意度等 11 个指标[2]。构建的河流健康评价指标体系结构图如表 1 所示。河流健康评价各指标级别和划分标准见表 2。

表 1 河流健康评价指标体系结构图

目标层	控制层	准则层	指标层
河流健康综合指数	自然环境	形态特征	河岸稳定性 x_1
		水文特征	河流流动性 x_2
			生态流量满足程度 x_3
		环境特征	水质综合指数 x_4
		生态特征	岸坡植被结构完整性 x_5
		灾害调节能力	河流生物多样性 x_6
		河流管理能力	防洪工程达标率 x_7
	社会服务	公众评价	岸线利用管理 x_8
		供水能力	公众满意度 x_9
			供水水量保证率 x_{10}
			水功能区水质达标率 x_{11}

由于河流健康评价系统的评价因素较多、层次结构复杂，各个指标的应用范围较广，适用性强，这些因素导致了模型评价的不确定性，而模糊识别评价方法可以有效地降低不确定因素对评价结果的影响，客观地描述评价系统的层次性和模糊性[3]。模糊识别评价法已被广泛运用到河流健康评价领域，任财[4]等人通过二元比较法确定指标权重，再用可变的模糊识别法建立评价模型对黄河下游河道健康进行评价并取得良好的实验效果；王笑宇[5]等通过赋权法与相对隶属度相结合的方式计算各指标权重，构建了贝叶斯与模糊识别相耦合的评价模型；刘营[6]等人通过 AHP 计算出指标权重并与指标隶属度矩阵进行耦合运算，对湛渎港河流健康状况进行了客观地评价，并提出一系列河流保护的建议。

本文建立了河流健康评价指标的隶属度函数矩阵，通过随机森林 Gini 指数计算出各个指标的权重，最后将两者进行耦合计算形成最终的模糊评价结果，实验结果验证了该方法的合理性与优越性，为某市河道的管理与整治起到了一定的作用。

2 改进的模糊识别健康评价方法

2.1 数据的预处理

模型构建的过程中，需对样本数据进行归一化的预处理，统一每个指标的变化范围，降低特征之间的差异性，利于随机森林计算各个评价指标之间的权重。设样本集为 $R = \{x_{ij} \mid i = 1,2,\cdots,n; j = 1,2,\cdots p\}$，$i,j$ 分别为样本序号和指标序号，x_{ij} 为样本值，利用如下公式对样本指标集进行归一化处理。若指标值越大分级越优：

$$x_{ij}^{*} = \frac{x_{maxj} - x_{ij}}{x_{maxj} - x_{minj}} \tag{1}$$

若指标值越小分级越优：

$$x_{ij}^{*} = \frac{x_{ij} - x_{minj}}{x_{maxj} - x_{minj}} \tag{2}$$

其中，x_{maxj}，x_{minj} 分别为特征指标 j 的最大值和最小值，x_{ij}^{*} 为特征指标归一化后的值。

2.2 随机森林(Gini 指数)确立指标权重

决策树是数据挖掘领域一种比较典型的单分类器，可以把它看作一个树形结构的模型，通过典型的节点展现树的特征，分别为：根节点、中间节点、叶子节点。决策树从根节点出发，再经过许多个中间节点，最后到达叶子节点，输出单一值，即每棵决策树到达唯一的叶子节点，实现了数据集的分类。为了解决决策树分类规则复杂、易得到局部最优解、过度拟合等问题，集成单个分类器，这就是随机森林[7]的思想。

表 2　河流健康评价各指标级别和划分标准

评价指标	指标级别及其划分标准			
	优	良	中	差
x_1	[0.85,1]	[0.70,0.85]	[0.40,0.70]	[0,0.40]
x_2	[0.80,1]	[0.60,0.80]	[0.40,0.60]	[0,0.40]
x_3	[0.95,1]	[0.90,0.95]	[0.80,0.90]	[0,0.80]
x_4	[4,5]	[3,4]	[2,3]	[0,2]
x_5	[0.85,1]	[0.60,0.85]	[0.40,0.60]	[0,0.40]
x_6	≥3.0	[2.0,3.0)	[1.0,2.0)	[0,1.0]
x_7	[95,100]	[85,95]	[65,85]	[0,65]
x_8	[0.90,1]	[0.70,0.90]	[0.50,0.70]	[0,0.50]
x_9	[90,100]	[70,90]	[45,70]	[0,45]
x_{10}	[95,100]	[85,95]	[65,85]	[0,65]
x_{11}	[80,100]	[70,80]	[50,70]	[0,50]

随机森林在特征随机选取后，需要通过节点分裂算法进行最优属性的选取，且采用程序递归的方式，将根节点分为两颗子树，又从选中的子树继续生成左右子树，如此递归，直到生成最终的叶子节点。节点分裂算法有很多种，包括 ID3、C4.5、CART 等。本文主要使用 CART 算法，它采用的分裂方式是 Gini 指标最小原则，Gini 指标是衡量特征属性重要度的方式。

假设经过数据预处理的集合 R^* 中的特征指标对应的样本 $R^{\#}$ 中包含 J 个类别，则其基尼指数为：

$$gini(T) = 1 - \sum_{j=1}^{N} p_j^2 \tag{3}$$

其中，p_j 为第 j 类样本的概率，在一次分割后集合 R^* 分成了 m 个部分 $\{N_1, N_2, \cdots, N_m\}$，则分割的基尼指数 $gini_{split}(T)$ 为：

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \cdots + \frac{N_m}{N} gini(T_m) \tag{4}$$

最终的 $gini_{split}(T)$ 作为每个特征样本对应的基尼指数，则各个特征指标基尼指数集合设为 $G = \{g_1, g_2, \cdots, g_j\}$，每个特征指标对应的权重为：

$$\theta_j = \frac{g_j}{\sum_{i=1}^{j} g_i} \tag{5}$$

最终得到评价指标的权重集合 $\theta = \{\theta_1, \theta_2, \cdots, \theta_j\}$。

2.3　改进的模糊综合评价模型

河流健康评价指标体系是一个多个因素的集合，根据评价指标及其评价等级，建立健康评价指标的隶属度函数 $W_{ij} = \begin{Bmatrix} w_{11} & w_{12} & \cdots & w_{1i} \\ w_{21} & w_{22} & \cdots & w_{2i} \\ \cdots & \cdots & \cdots & \cdots \\ w_{j1} & w_{j2} & \cdots & w_{ji} \end{Bmatrix}$，$i$ 表示健康评价指标序号，j 表示评价指标等级标准序号。

相对隶属度在模糊集合论中表述一类过渡或中介，可以对模糊性概念进行精确表述[8-9]。通过规定评价指标等级 j 的健康程度确定评价指标 i 的相对隶属度，当 $j = 1$ 时，表示健康程度为差，评价指标 i 的标准值 y_{i1} 对健康程度的相对隶属度为 $s_{j1} = 0$；当 $j = c$ 时，表示健康程度为优，评价指标 i 的标准值 y_{ic} 对健康程度的相对隶属度为 $s_{j1} = 1$；则当 $j = i$ 时，评价指标 i 的标准值 y_{ji} 对健康程度的相对隶属度则为：

$$s_{ji} = \frac{y_{ji} - y_{j1}}{y_{jc} - y_{j1}} \tag{6}$$

同理，评价指标的实测值 x_{jk} 对应的健康程度转化为对应的相对隶属度 r_{jk}：

$$r_{jk} = \begin{cases} 1 & x_{jk} > y_{jc} \\ (x_{jk} - y_{j1})/(y_{jc} - y_{j1}) & y_{j1} < x_{jk} < y_{jc} \\ 0 & x_{jk} < y_{j1} \end{cases} \tag{7}$$

算法详细步骤如下：

步骤 1　根据某地河流所处区域的综合生态情况，分别从社会和自然两个角度出发，在目标层、控制层、准则层、指标层 4 个层次的基础上构建河流

健康评价指标。

步骤 2 设实际测量的河流健康指标的样本集为 $R = \{x_{ij} \mid i = 1,2,\cdots,n;j = 1,2,\cdots p\}$，$i,j$ 分别为样本序号和指标序号，x_{ij} 为样本值，根据实测样本集每个特征的进行归一化等数据预处理，得到新的样本集 $N = \{x_{ij}^{*} \mid i = 1,2,\cdots,n;j = 1,2,\cdots p\}$，$i,j$ 分别为样本序号和指标序号，x_{ij}^{*} 为样本值。

步骤 3 对步骤 1 中的新的样本集 N 训练生成 RF 模型，根据 *gini* 指数选取重要特征，计算每个特征的 *gini* 指数，得到每个评价指标的权重为 $\theta = \{\theta_1,\theta_2,\cdots,\theta_j\}$。

步骤 4 根据实测评价指标样本集 R，建立健康评价指标的隶属度函数 $W_{ij} = \begin{Bmatrix} w_{11} & w_{12} & \cdots & w_{1i} \\ w_{21} & w_{22} & \cdots & w_{2i} \\ \cdots & \cdots & \cdots & \cdots \\ w_{j1} & w_{j2} & \cdots & w_{ji} \end{Bmatrix}$，$i$ 表示健康评价指标序号，j 表示评价指标等级标准序号。

步骤 5 将实测样本集的隶属度矩阵 W_{ij} 及各指标的 *gini* 权重集 θ 进行耦合运算 $Z = \theta \times W$，得到模糊综合评价结果 $Z = \{Z_1,Z_2,\cdots,Z_i\}$，选取 Z 中的最大值 $Z_{\max}$ 作为最终的模糊综合评价结果[10]。

3 实验分析

2017～2018 年期间，课题组在某市水利部门的委托下对该市市区内 6 条主要河流健康状况进行考察，根据实地勘查数据和水利部门提供的资料，将 6 条河流样本集的实测数据列于表 3 中。

由表 3 中的健康评价指标实测数据，根据公式(6)、(7)计算出各个指标的隶属度矩阵：

$$R = \begin{Bmatrix} 1.00 & 0.93 & 0.96 & 1.00 & 0.98 & 0.85 \\ 0.76 & 0.71 & 0.73 & 0.68 & 0.56 & 0.29 \\ 0.98 & 0.96 & 0.97 & 1.00 & 0.98 & 0.99 \\ 0.66 & 0.65 & 0.65 & 0.66 & 0.56 & 0.53 \\ 0.76 & 0.87 & 0.84 & 0.85 & 0.79 & 0.80 \\ 0.34 & 0.35 & 0.34 & 0.47 & 0.46 & 0.42 \\ 0.98 & 0.89 & 0.88 & 0.99 & 0.87 & 0.76 \\ 0.97 & 0.94 & 0.92 & 1.00 & 0.91 & 0.88 \\ 0.91 & 0.96 & 0.90 & 0.91 & 0.84 & 0.83 \\ 0.98 & 1.00 & 0.96 & 0.99 & 1.00 & 0.98 \\ 0.55 & 0.54 & 0.58 & 0.53 & 0.41 & 0.40 \end{Bmatrix}$$

对上述健康评价指标实测数据进行归一化处理，并计算每个特征的 *gini* 指数，如图 1 所示。

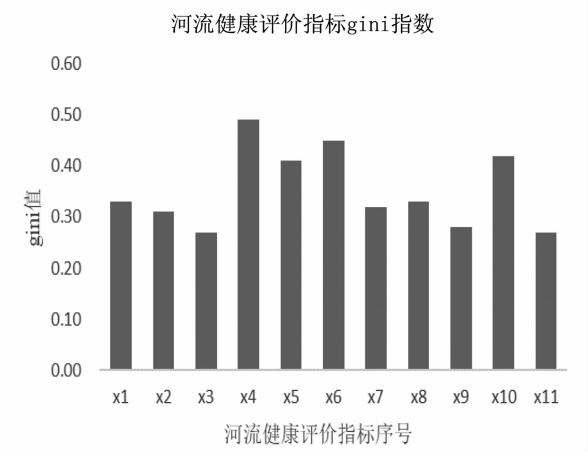

图 1 河流健康评价指标各指标 *gini* 指数

最终根据上述各个评价指标的 *gini* 值，可计算

表 3 6 条河流样本集健康评价指标实测值

指标	河流编号					
	1	2	3	4	5	6
x_1	0.89	0.79	0.82	0.91	0.83	0.72
x_2	0.61	0.57	0.58	0.54	0.45	0.23
x_3	0.93	0.91	0.92	0.95	0.93	0.94
x_4	2.62	2.59	2.61	2.64	2.23	2.12
x_5	0.65	0.74	0.71	0.72	0.67	0.68
x_6	1.01	1.04	1.02	1.42	1.39	1.27
x_7	0.93	0.85	0.84	0.94	0.83	0.72
x_8	0.87	0.85	0.83	0.91	0.82	0.79
x_9	82	86	81	82	76	75
x_{10}	93	95	91	94	95	93
x_{11}	44	43	46	42	33	32

求得各指标的权重比例 $\theta = \{0.085, 0.080, 0.070, 0.126, 0.106, 0.116, 0.082, 0.085, 0.072, 0.108, 0.070\}$。由矩阵 R 和权重 θ 计算 $Z = \theta \times W = \theta = \{0.085, 0.080, 0.070, 0.126, 0.106, 0.116, 0.082, 0.085, 0.072, 0.108, 0.070\}$

$$\begin{Bmatrix} 1.00 & 0.93 & 0.96 & 1.00 & 0.98 & 0.85 \\ 0.76 & 0.71 & 0.73 & 0.68 & 0.56 & 0.29 \\ 0.98 & 0.96 & 0.97 & 1.00 & 0.98 & 0.99 \\ 0.66 & 0.65 & 0.65 & 0.66 & 0.56 & 0.53 \\ 0.76 & 0.87 & 0.84 & 0.85 & 0.79 & 0.80 \\ 0.34 & 0.35 & 0.34 & 0.47 & 0.46 & 0.42 \\ 0.98 & 0.89 & 0.88 & 0.99 & 0.87 & 0.76 \\ 0.97 & 0.94 & 0.92 & 1.00 & 0.91 & 0.88 \\ 0.91 & 0.96 & 0.90 & 0.91 & 0.84 & 0.83 \\ 0.98 & 1.00 & 0.96 & 0.99 & 1.00 & 0.98 \\ 0.55 & 0.54 & 0.58 & 0.53 & 0.41 & 0.40 \end{Bmatrix}$$

$= \{0.790, 0.785, 0.777, 0.812, 0.751, 0.696\}$。由表 4 可以看出，6 条河流的各项健康评价指标等级经过随机森林与模糊识别相结合的方法改进后精度较改进前的层次分析模糊识别方法有一定的提高，预测率达 67%。编号为 1 ~ 5 的河流评价结果都在[0.70,0.85]之间，评价等级为良；编号为 6 的河流评价结果在 0.70 以下，评价等级为差，说明该市河流的整体健康状况良好，6 条河流均能满足市区居民的实际用水需求，但是仍然存在改进的空间，希望当地相关部门加强对市区河流的治理与监管，改善水质条件，提高河流生物多样性，加强河流稳定性建设，维护生态系统的稳定性。综合上述分析结果，本文结果合理、可行，与实际情况相符。

表 4　河流健康评价结果

河流编号	1	2	3	4	5	6
本文方法	0.790（良）	0.785（良）	0.777（良）	0.812（良）	0.751（良）	0.696（中）
AHP 模糊	0.785	0.780	0.773	0.805	0.743	0.686
是否一致	√	√	√	√	×	×

4　结论

1）河流健康评价是一个多指标评价过程，本文构建的随机森林和模糊识别耦合模型有一定的适用性。文中提出的相对隶属度函数替代了传统的求权值方法，提高了模型的评价精度，降低了不确定性因素对模型的影响。

2）本文通过随机森林 *gini* 指标计算各个健康评价指标的权重避免了层次分析法中专家打分的主观性及局限性，具有一定的参考价值；且当数据集中的样本量越大，所计算的指标权重越合理，耦合模型的评价精度越高。

参考文献：

[1] 刘存，徐嘉，张俊，等．国内河流健康研究综述[J]．海河水利，2018(04)：6－12．

[2] 吕照根，周必翠，舒持恺，等．河流健康评价指标体系合理性研究分析[J]．江苏水利，2017(09)：10－14．

[3] 崔嘉宇，张宁红，郁建桥，等．改进的模糊评价法在太湖水质评价中的应用[J]．环境工程学报，2015，9(11)：5357－5363．

[4] 任财，陈守煜，周惠成．基于可变模糊识别的黄河下游河道健康评价[J]．人民黄河，2014，36(02)：45－48．

[5] 王笑宇，王国玖，李娜，等．贝叶斯公式与模糊识别耦合方法在河流健康评价中的应用[J]．水电能源科学，2017，35(01)：48－52．

[6] 刘营．河流健康的模糊评价模型研究[J]．广西水利水电，2018(02)：95－99．

[7] 曹正凤．随机森林算法优化研究[D]．北京：首都经济贸易大学，2014．

[8] 张彦霞，肖清泰，肖汉杰，等．组合权重优化的企业财务绩效可变模糊综合评价模型[J]．数学的实践与认识，2018，48(03)：64－74．

[9] 方运海，郑西来，彭辉，等．基于模糊综合与可变模糊集耦合的地下水质量评价[J]．环境科学学报，2018，38(02)：546－552．

[10] 彭鹏菲，刘忠，张建强．基于评价指标模糊集的复杂系统效能评估方法研究[J]．指挥控制与仿真，2007(03)：58－61．