基于 SVM-RF-ANN 集成模型的珠江流域水质预测

李伟奇，黄炫，朱正棋
广州大学 经济与统计学院，广东 广州 510006

DOI:10.61369/ASDS.2025100012

摘　　要： 水资源短缺已成为制约区域可持续发展的关键性自然资源瓶颈。作为高度城市化区域，珠江流域面临严峻的水质性缺水问题，实现准确的水质预测是保障水生态安全的重要前提。传统水质预测方法多依赖专家经验，存在效率低、误差累积显著等局限性。本文基于机器学习方法，收集2015–2023年珠江流域部分代表性断面的水质监测数据，并进行规范化预处理。采用支持向量机、随机森林与人工神经网络构建水质类别分类模型，运用网格搜索与交叉验证方法优化模型参数，结合特征重要性分析量化各水质指标对分类结果的贡献度。进一步引入软投票集成策略，进而采用粒子群优化 (PSO) 算法动态确定最优权重，融合各基模型优势以提升整体预测性能。研究结果可为珠江流域水环境精细化管理提供科学依据，有助于提升水质预测的准确性与鲁棒性，推动流域水资源的可持续利用与生态保护。

关 键 词： 机器学习；水质预测；珠江流域；SVM；RF；ANN

Water Quality Prediction in the Pearl River Basin Using an SVM-RF-ANN Ensemble Model

Li Weiqi, Huang Xuan, Zhu Zhengqi
School of Economics and Statistics, Guangzhou University, Guangzhou, Guangdong 510006

Abstract: Water scarcity presents a critical bottleneck to regional sustainable development. The highly urbanized Pearl River Basin suffers from acute water quality–induced scarcity, making accurate prediction essential for aquatic ecological security. Traditional methods, largely reliant on expert experience, are often inefficient and prone to error accumulation.This study develops a machine learning framework for water quality classification using monitoring data from 2015 to 2023 in the Pearl River Basin. Three algorithms—Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN)—were applied, with hyperparameters optimized via grid search and cross–validation. Feature importance analysis was conducted to identify key water quality indicators.A weighted soft–voting ensemble was implemented, with weights dynamically optimized using Particle Swarm Optimization (PSO), effectively integrating the advantages of each base model. The proposed approach demonstrates enhanced overall accuracy and robustness, offering a reliable tool for refined water quality management and supporting the sustainable utilization and conservation of water resources in the basin.

Keywords: machine learning; water quality prediction; Pearl River Basin; SVM; RF; ANN

引言

珠江流域水质研究兼具重要的生态安全、经济发展、政策制定与科研创新价值。在生态层面，水质污染直接威胁水生态系统平衡，影响鱼类产卵与栖息环境[1]；在饮水安全方面，部分水源地水质不达标与中小河流污染问题依然突出[2]；经济发展层面，2024年广东省生态环境公报显示总磷、化学需氧量与氨氮仍为主要超标指标，由此引发的污染问题已对农业、工业与旅游业产生连锁影响[3]；政策层面，精准水质预测可为枯水年水量分配与污水处理设施规划提供关键技术支撑[4]。在科研层面，该领域已呈现出多技术融合的研究态势，例丁号楠运用多元统计解析水质时空特征[5]，白雯睿等构建 VMD-CNN-LSTM 预测模型[6]，牛亚朝等探索珠三角水资源生态足迹[7]，这些工作推动了环境科学与数据科学的交叉创新。

基金项目：2024年度广东省自然科学基金面上项目“基于判别与深度表示的高维数据聚类方法研究”（2024A1515012040）。

作者简介：
李伟奇，广州大学经济与统计学院，硕士研究生，研究方向为机器学习与计算机视觉；
黄炫，广州大学经济与统计学院，硕士研究生，研究方向为机器学习与计算机视觉；
朱正棋，广州大学经济与统计学院，硕士研究生，研究方向为机器学习与计算机视觉。

当前，水质预测研究持续深化。国内学者发展了多元统计、VMD-CNN-LSTM 等多种方法[5,6]，国外则注重卫星与传感器数据融合及复杂人工智能技术[8,9]。然而，现有方法在应对极端气候事件、复杂污染机制及模型实时性方面仍存在明显不足。

本文集成随机森林、支持向量机与人工神经网络三种机器学习算法，引入粒子群优化算法动态确定最优权重，构建加权组合预测模型，旨在提升珠江流域水质预测的准确性与鲁棒性，为流域水环境精细化管理提供新方法。

一、研究区域数据来源及预处理

（一）数据来源

本文以珠江流域内蕉门、流溪河、增江口等重要国家地表水考核断面为研究对象[5]，水质监测数据来源于水利部珠江水利委员会 (www.pearlwater.gov.cn)。所用指标包括电导率 (EC)、pH、溶解氧 (DO)、高锰酸盐指数 (CODMn)、化学需氧量、生化需氧量等。

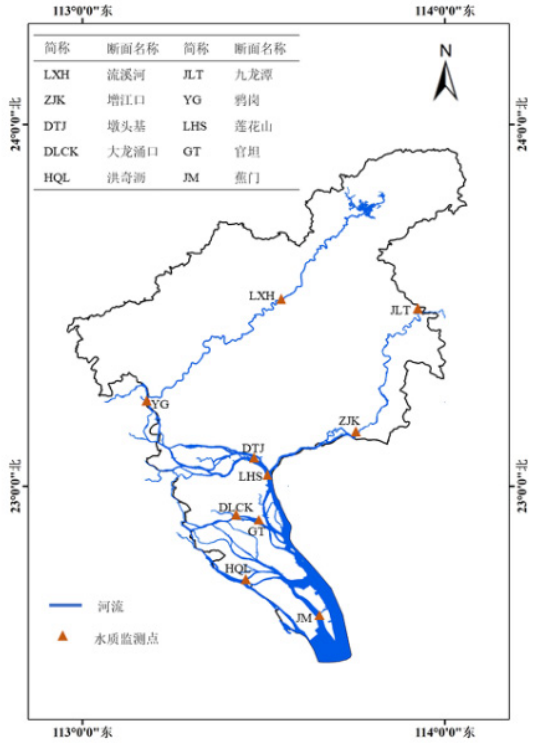

图1：部分水质监测站点位置图

（二）数据预处理

针对原始数据存在的缺失、格式不一与量纲差异问题，采用以下流程进行系统预处理：

1. 缺失值处理

针对缺失数据，将表示未监测的“-1”标记替换为 NaN，对缺失率超过70% 的指标予以剔除，其余数值型缺失值采用中位数填充，以保持数据分布的稳健性。

2. 类别变量数值化

将文本格式的“综合水质类别”（如“Ⅰ类”至“劣Ⅴ类”）进行数值编码映射为0－5的离散标签，以适应模型输入要求。

3. 数据划分与标准化

将完整数据集按7:3的比例划分为训练集与测试集，划分过程设置固定随机种子以确保结果可重现。采用 Z-score 标准化方法对特征数据进行归一化处理，消除量纲影响，提升模型训练效率与预测精度，其公式为：

$$x_{标准化} = \frac{x - \mu}{\sigma} \quad (1\text{-}1)$$

其中，x 代表原始数据，μ 是特征的均值，θ 是特征的标准差。

二、理论背景和相关工作

本文采用随机森林（RF）、支持向量机（SVM）和人工神经网络（ANN）三种算法，并构建一种基于粒子群优化（PSO）的加权组合模型，以应对珠江流域水质预测的复杂性。

（一）随机森林的思想及基本原理

随机森林属于 Bagging 类集成算法，其核心机制通过 Bootstrap 抽样构建决策树基学习器，并在节点分裂时引入特征随机选择，以此增强模型多样性并提升泛化性能。对于回归类任务，随机森林的最终输出为所有决策树预测值的算术平均，其数学表达为：

$$y^*(x) = \frac{1}{N}\sum_{i=1}^{N} y_i^*(x) \quad (2\text{-}1)$$

其中N为森林中树的数量，$y^*(x)\,(i=1,2,...,N)$为每棵树对样本x的预测值。

杨宇锋等[10]应用该模型成功实现了辽河流域氮、磷浓度的高精度预测，验证了其在水质预测任务中的有效性。本文将其引入珠江流域水质分类，旨在利用其集成优势捕捉多水质指标与综合类别间的复杂非线性映射。

（二）支持向量机的思想及基本原理

支持向量机（Support Vector Machine, SVM）是一种基于统计学习理论的监督学习模型，其优化目标为在特征空间中构造最大间隔分类超平面，并通过核函数映射处理线性不可分问题。

设有线性可分的训练样本集$\{(x_i,y_i)\}_{i=1}^{n}$，其中$y_i\in\{-1,+1\}$为类别标签，SVM 的目标是求解如下凸优化问题：

$$\min_{w,b} \frac{1}{2}|w|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \quad \forall i \qquad (2\text{-}2)$$

其中，w 为法向量，b 为偏置项。该问题的对偶形式可通过引入拉格朗日乘子进行推导，并利用核函数将原始特征映射到高维空间，以处理线性不可分的情形。常用的核函数包括线性核、多项式核与径向基核等。

丁彦蕊等 [11] 采用 SVM 对太湖入湖河流水质关键影响因素进行辨识，证实了该方法在水质特征分析与分类中的有效性。本文借助其结构风险最小化特性，构建珠江流域水质类别的最优判别边界，以提升小样本条件下的分类稳定性。

（三）人工神经网络的思想及基本原理

人工神经网络通过模拟生物神经元连接机制构建多层感知机模型，以前向传播实现信号传递，并以误差反向传播算法进行参数优化。

网络中的每个神经元接收前一层神经元的输出，并通过加权求和与非线性激活函数（如 Sigmoid、ReLU 等）产生其输出值。第 l 层第 j 个神经元的输出 $a_i^{(l)}$ 可表示为：

$$a_j^{(l)} = f\left(\sum_i w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right) \qquad (2\text{-}3)$$

其中，$w_{ji}^{(l)}$ 表示连接第 $l-1$ 层第 i 个神经元与第 l 层第 j 个神经元的权重，$b_j^{(l)}$ 为对应的偏置项，$f(\cdot)$ 为非线性激活函数。

赵颖等 [12] 将改进的人工神经网络成功应用于河南部分河流的水质评价，验证了该方法在水质分类任务中的有效性与适用性。本研究利用其强大的非线性拟合能力，模拟珠江流域咸潮入侵等复杂动态水文过程。

（四）组合模型

为集成单一模型的优势并提升整体预测性能，本文构建了一种基于 PSO 的加权集成模型。该模型通过粒子群优化算法动态确定各基模型的最优权重系数，以实现偏差 – 方差权衡。

郭建青等 [13] 将粒子群优化算法应用于河流水质参数优化，验证了该优化方法在水环境研究中的有效性。本文采用 PSO 算法优化各基模型的权重系数，构建如下集成表达式：

$$\hat{y}_{\text{ensemble}} = \omega_{\text{RF}} \cdot \hat{y}_{\text{RF}} + \omega_{\text{SVM}} \cdot \hat{y}_{\text{SVM}} + \omega_{\text{ANN}} \cdot \hat{y}_{\text{ANN}} \qquad (2\text{-}4)$$

其中，$\hat{y}_{\text{RF}}$，$\hat{y}_{\text{SVM}}$，$\hat{y}_{\text{ANN}}$ 分别代表三个模型的预测值；ω_{RF}，ω_{SVM}，ω_{ANN} 为经由 PSO 算法优化得到的非负权重系数，并满足如下约束条件：

$$\omega_{\text{RF}} + \omega_{\text{SVM}} + \omega_{\text{ANN}} = 1 \qquad (2\text{-}5)$$

该集成策略旨在自适应地融合各基模型的优势，从而在面对珠江流域复杂的非线性水质动态时，展现出更强的鲁棒性与泛化能力。

三、实验结果

本文基于2015–2024年珠江流域10年重要流域断面的水质监测数据展开，以2015–2021年数据作为训练集，2022–2024年3年数据作为测试集，对随机森林（RF）、支持向量机（SVM）、人工神经网络（ANN）及 PSO 加权集成模型进行了综合对比分析。

（一）模型性能评估

使用准确率、精确率、召回率、F1分数、类别1精确率等指标作为模型的性能评估指标，对各个模型进行了从不同方面的评估（表1）。

表1：模型评价对比

模型	准确率	精确率	召回率	F1分数	类别1精确率	类别1召回率	抗噪声准确率
SVM	0.76	0.74	0.76	0.75	0.81	0.90	0.76
Random Forest	0.93	0.95	0.90	0.97	0.97	0.91	0.87
ANN	0.81	0.82	0.81	0.81	0.86	0.90	0.78
组合模型	0.94	0.94	0.94	0.93	0.99	0.96	0.89

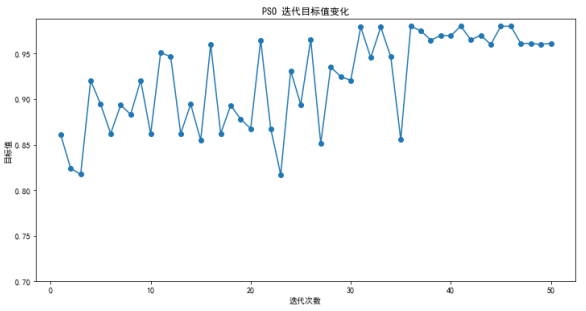

图2：PSO 迭代目标值变化图

由表1可知，随机森林在多项评估指标中表现最优，其准确率达到0.93，F1分数高达0.97，且具备良好的抗噪性能。支持向量机与人工神经网络的准确率分别为0.76与0.81；其中 SVM 训练耗时较长，ANN 对输入噪声较为敏感，当引入10% 异常值时，其准确率下降至0.78。基于 PSO 加权的集成模型综合性能最佳，准确率提升至0.94，F1分数为0.93，在咸潮入侵等复杂水文情景中预测误差显著降低，展现出更优的泛化能力。

同时，各个模型的混淆矩阵（图3至图6）显示，在各类别预测中，随机森林模型的混淆矩阵对角线上具有更高的数值，表明其分类准确率更高且误判更少。SVM 模型在混淆矩阵中的非对角线数值较大，表明其在各个类别上存在一定的误判。组合模型（PSO 加权重）的混淆矩阵中，在各类别上数值准确，组合模型在各类别上均具有较高的预测准确度。

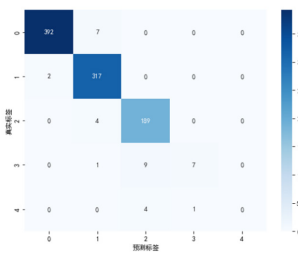

图3：RF 模型混淆矩阵

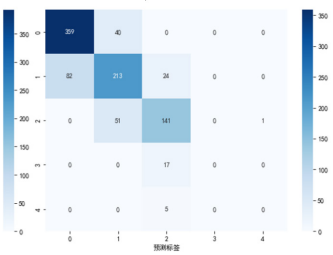

图4：SVM 模型混淆矩阵

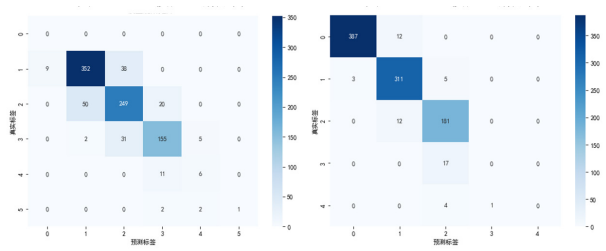

图5：ANN 模型混淆矩阵　图6：组合模型（PSO 加权）混淆矩阵

（二）模型局限性及改进方向

本文基于机器学习方法构建的水质预测模型在珠江流域应用中表现出较为良好的预测性能，但仍存在以下可改进的方向：

1. 数据维度局限性：上游监测站点稀疏区域的预测误差达18%，后续研究可结合灰色预测法[14]或迁移学习[15]方法，提升小样本数据区域的预测稳定性。

2. 过程机制融合：当前模型对污染物迁移转化过程的物理机制刻画不足，后续可引入流速、流量等水动力参数[16]，构建机理与数据驱动融合的混合模型。

3. 实时预测优化：针对模型训练耗时较长的问题，可研究模型轻量化技术或增量学习算法，以降低计算成本，提升预测时效性。

四、结论

本文构建了基于 SVM-RF-ANN 的机器学习模型体系，实现了对珠江流域水质类别的有效预测。其中，随机森林在特征可解释性方面表现突出，特征重要性分析显示氨氮、总磷等营养盐指标对水质类别的贡献率超过50%，为核心驱动因子。通过引入粒子群优化算法确定最优权重，建立的加权集成模型显著提升了预测性能，准确率达到94.0%，F1分数为0.93。该模型在咸潮入侵等复杂水文条件下仍保持较强鲁棒性，验证了集成学习策略在应对非线性水质预测问题中的有效性。基于模型分析结果，珠江流域水质管理需重点关注中下游工业点源与农业面源污染，并加强枯水期生态基流保障，以提升水体自净能力。

未来研究将着力于以下方向：一是构建融合水动力学机理与机器学习的“物理机制—数据驱动”混合模型，以弥补现有模型对污染物迁移转化过程刻画的不足；二是整合遥感、无人机与物联网等多源监测数据，提升对新型污染物的识别预测能力；三是引入不确定性量化研究，为应急决策提供风险量化依据；四是结合数字孪生技术，开发珠江流域水质预测与调度一体化平台，实现跨省界断面实时预警、污染源追溯及生态补偿政策效果模拟，支撑粤港澳大湾区 “水－态－经济” 协同发展。

参考文献

[1] 王建平 . 珠江流域生态水力学研究进展 [J]. 中国水利，2023,（14）: 34-38.

[2] 罗昊，周雪欣 . 珠江流域饮用水水源保护现状及对策 [J]. 水利技术监督，2024,（08）: 70-72.

[3] 广东省生态环境厅 . 2024 年广东省生态环境状况公报 [R/OL]. (2025-06-05) [2025-10-10]. 2024 年广东省生态环境状况公报 - 广东省生态环境厅公众网

[4] 郑冬燕 . 珠江流域水量分配基本框架研究 [C]// 中国水利学会2013学术年会论文集——S1 水资源与水生态，2013: 444-447.

[5] 丁号楠 . 珠江（广州段）流域河流水质时空特征分析和预测研究 [D]. 华南理工大学，2023.

[6] 白雯睿，杨毅强，郭辉，等 . 基于 VMD-CNN-LSTM 的珠江流域水质多步预测模型研究 [J]. 四川轻化工大学学报（自然科学版），2022, 35(04): 66-74.

[7] 牛亚朝，罗柱，王强，等 . 珠三角区域水资源生态足迹动态分析与预测 [J]. 人民珠江，2024, 45(05): 34-45.

[8]van Vliet, M.T.H., Thorslund, J., Strokal, M. et al. Global river water quality under climate change and hydroclimatic extremes. Nat Rev Earth Environ 4, 687 - 702 (2023).

[9]Wang, M., Bodirsky, B.L., Rijneveld, R. et al. A triple increase in global river basins with water scarcity due to future pollution. Nat Commun 15, 880 (2024).

[10] 杨宇锋，武暕，王璐，等 . 基于随机森林模型的辽河高时间分辨率氮、磷浓度模拟与预测 [J]. 环境科学学报，2022,42(12):384-391

[11] 丁彦蕊，孙小妹，王文超，等 . 基于支持向量机的太湖入湖河流水质影响因素的研究 [J]. 水资源与水工程学报，2011,22(05):38-40+46.

[12] 赵颖，王建英，孙燕，等 . 改进人工神经网络在河南部分河流的水质评价中的应用 [J]. 环境与发展，2018,30(03):216-217.

[13] 郭建青，李彦，王洪胜，等 . 粒子群优化算法在确定河流水质参数中的应用 [J]. 水利水电科技进展，2007,(06):1-5.

[14] 李娜，王腊春，谢刚，等 . 山东省辖淮河流域河流水质趋势的灰色预测 [J]. 环境科学与技术，2012,35(02):195-199.

[15] 杨晶，路恒通，金鑫 . 机器学习赋能智慧水利的应用现状 [J]. 水利水电技术，2024,55(10):137-147.

[16] 聂影，刘永宏，梁卫芳，等 . 基于物联网 + 机器学习的水位、水质预测模型应用研究 [J]. 物联网技术，2024, 14(10): 89-94.

