文章编号：1672 - 3317（2023）11 - 0112 - 09

灌区用水调度的知识图谱模型构建

——以泾史杭灌区瓦西干渠灌域为例

苏 楠[1]，章少辉[1,2*]，白美健[1,2]，张宝忠[1,2]
（1.中国水利水电科学研究院，北京 100038；2.流域水循环模拟与调控国家重点实验室，北京 100038）

摘 要：【目的】人工智能技术以全新的自适应学习视角为灌区智慧化建设带来了异于传统的研究范式，迄今已有众多智能算法与技术应用于灌区用水调度决策研究中，但已有成果缺乏便利性和易操作性，无法真正落地。【方法】为此，本文通过对比分析人工智能技术中的随机森林和 BP 神经网络模型，发现融合 SHAP（Shapley additive explanation，SHAP）方法的随机森林模型具有更优的灌区用水调度预测效果。在此基础上，基于历史实测数据将灌区用水调度变量划分为不同梯度，表征出 9.82×10^{11} 组用水调度场景样本，并采用预先训练的随机森林模型获得了各样本对应的调度流量预测值，由此形成了灌区用水调度场景与调度流量值之间映射关系的基础数据库，即在现有调度历史数据的基础上利用机器学习模型丰富和加密了调度场景，得到了能实现基本覆盖现实调度场景的调度场景库。基于该场景数据库，利用 Neo4j 图形库构建出泾史杭灌区用水调度流量预测值知识图谱模型。【结果】利用该图谱模型，灌区用水调度管理人员仅需确认目标调度场景中各调度变量在知识图谱模型中的近似梯度值，即可检索获得调度流量预测值。【结论】经应用验证表明，由该知识图谱模型获得的调度流量预测值误差在泾史杭灌区用水管理人员的经验认知范围内，且可实现调度流量值的实时检索。

关 键 词：泾史杭灌区；用水调度；机器学习；SHAP 方法；知识图谱
中图分类号：S274 **文献标志码：**A doi：10.13522/j.cnki.ggps.2023047 OSID：

苏楠，章少辉，白美健，等. 灌区用水调度的知识图谱模型构建：以泾史杭灌区瓦西干渠灌域为例[J]. 灌溉排水学报，2023, 42(11): 112-120.
SU Nan, ZHANG Shaohui, BAI Meijian, et al. Construction of Knowledge Graph Model for Irrigation Water Scheduling[J]. Journal of Irrigation and Drainage, 2023, 42(11): 112-120.

0 引 言

【研究意义】灌区作为国家水网建设的重要内容，其合理的用水调度对于工程的高效安全运行和水资源可持续利用，具有重要的现实意义。对灌区管理人员而言，用水调度的目标变量通常是总闸的入流量过程，与之相关的变量涉及其控制范围内的需水单元布局、生态环境特征、降水人文境况等诸多方面，具有定性和定量双层面的典型属性。这就导致经典的、能够精细刻画自然物理过程的数理模型难以充分发挥作用，故迄今仍缺少能与用水调度人员历史经验相匹配的、切实可行的灌区用水调度模型。为此，人们开始寻求人工智能等新技术与方法。

【研究进展】人工智能技术中诸如随机森林模型、人工神经网络以及由此延伸发展的深度学习模型等，近年来在低成本高性能算力凸显的大环境下得到了充分发展并开始被广泛应用，其突出特点是，能有效应对具有定性和定量两种属性变量的复杂事件与过程，这与其具有复杂的网络结构、能够自适应学习和调整海量参数密切相关[1-3]。迄今，人工智能模型已然被应用于灌区用水调控的模拟计算中[4-5]。考虑到灌区用水调控是一个真实复杂的、具有时间演变特征的场景，也同样有将算法模型与其他工具相结合寻求更优方法的研究，例如 Arcgis、人工蚁群算法、随机森林等[6-8]。【切入点】但由于模型应用对使用者有一定技术要求不具便利性，以及需将模型作为底层结构依托于一定的应用界面开发成本较高等原因，当前多数灌区真实的用水调度过程，并未能将智能算法真正融合到实际调度应用中。特别是由于运行成本、维护更新费用等资金方面的问题，当前灌区信息化系统的覆盖范围受到很大程度的限制[9]。【拟解决的关键问题】故本文充分融合人工智能技术中各模型的特点，以此来集中应对灌区用水调度这个实际命题，以期找到能便利应用、将模型与实际应用紧密衔接且构建成

收稿日期：2023-02-16 修回日期：2023-08-10 网络出版日期：2023-11-13
基金项目：中国水科院专项项目（ID0145B052021）
作者简介：苏楠（1997-），女，云南楚雄人。硕士研究生，主要从事灌区用水管理研究。E-mail: sue_sn_email@163.com
通信作者：章少辉（1977-），男，河北石家庄人。教授级高级工程师，博士，主要从事水动力学模拟与灌区用水调控研究。
E-mail: zhangsh@iwhr.com

本低的方法，实现灌区用水调度中调度流量的预测。

本文拟对比分析随机森林模型和 BP 神经网络在非线性表征灌区用水调度目标变量与各自变量之间映射关系方面的精度，在此基础上，采用获得的最优映射关系形成大样本数据库，在现有历史数据基础上加密调度场景。进而采用知识图谱技术解决灌区用户与大样本数据库之间的实时交互问题，利用其对调度场景可进行检索查询的功能，建立切实可行的灌区用水调度分析工具。

1 研究区域及方法

1.1 研究区域

淠史杭灌区位于安徽六安，始建于 1958 年，包括淠河、史河和杭埠河 3 个子灌区，实际灌溉面积约 67 万 hm^2，是以防洪、灌溉、供水为主，兼有发电、生态、旅游等功能的综合性水利工程，是典型的引、蓄、提 3 种水源综合利用灌区[10]。本文关注的瓦西干渠位于淠河子灌区上段（图 1），渠首瓦西干渠进水闸的设计流量为 27.4 m^3/s，干渠长 61.2 km，其包括 15 条支渠，涉及灌溉面积约 3.84 万 hm^2。

1.2 数据来源

灌区用水调度的目标变量是流量，具体到本研究区域，则是瓦西干渠进水闸（后文简称瓦西进水闸）的实时调度流量（Q）。依据以往经验并结合淠史杭灌区实际调度能获取的特征变量类型，影响目标流量的因素（后文均称为特征变量）共有 8 类，如表 1 所示。特征变量中选择了罗管节制闸水位数据，因为其是总干渠上唯一的节制闸，其闸前水位需常年维持在 49~49.55 m 之间，以确保上游周边生活用水。表 1 所示的特征变量中，温度和降水数据具有显著的空间变异性，为定量表征局部小气候或水文水力特征，把各个观测点获得的实测数据单独作为特征变量，以反映其对目标变量的真实影响，由此形成 8 类共 73 个特征变量。

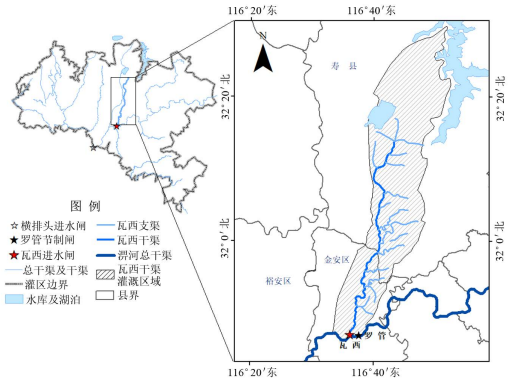


图 1 淠史杭灌区瓦西干渠灌域示意图

Fig.1 Schematic diagram of Washi canal irrigation area in Pishihang Irrigation District

表 1 特征变量及说明

Table 1 Characteristic variables and their description

因素	说明
用水调度流量 Q/(m^3 s^{-1})	目标变量，一动一记
60 个雨量站点测量的日降水量 $R_{1\text{-}60}$/mm	1~19 裕安区雨量站点，20~38 金安区雨量站点，39~60 寿县雨量站点
地区水利局和防办向灌区调度总局发出的调度请求 Y	YW 为瓦西闸，YL 为六冲测流站；其包含具体的流量要求及灌域需水量
城市和生态分阶段供水补水水量 C/万 m^3	
瓦西干渠周边中小型水库及塘坝蓄水 $S_{1\text{-}4}$/万 m^3	1 为中型水库，2 为小（一）型水库，3 为小（二）型水库，4 为塘坝；数据非灌期为 1 月一报，灌期为 1 旬一报
罗管节制闸水位 H/m	1 d 一记
灌区范围内 3 个气象站点记录的日气温数据 $T_{1\text{-}3}$/℃	
灌期和非灌期经验划分 G	1 为灌期（每年 4—9 月），0 为非灌期（10 月至次年 3 月）
时间 T/d	用 MMDD 格式表示

基于特征变量和目标变量这 2 类因果数据在时间上匹配的原则，收集淠史杭灌区瓦西干渠灌域内的上述变量数据，为建立目标变量和特征变量之间的非线性定量映射表征提供基础数据。其中，淠史杭灌区瓦西干渠进水闸用水调度流量记录、罗管节制闸水位数据、用水调度请求、中小型水库和塘坝蓄水量以及城市和生态分阶段供水补水水量数据均来自淠史杭灌区管理总局。日降水量数据来自六安市水文水资源局，如图 1 所示，从淠河渠首横排头到瓦西灌域，渠道共经过了六安市的裕安区、金安区和淮南市的寿县，为确保模型模拟的精度，搜集了这 3 个区县内共 60 个雨量站点的日降水数据，以考虑瓦西进水闸上游及干渠控制范围内降水对渠道流量的影响。逐日气温数据来自中国气象科学数据共享服务网的《中国地面气象资料数据集（V3.0）》。由于淠史杭灌区历史数据多采用手写形式记录储存，受数据记录详尽程度影响，本文选用的数据样本时间跨度为 2015—2020 年。

1.3 研究方法

1.3.1 不同调度场景的特征划分

灌区实际用水调度中，不同调度情景能获得的参考指标数据有差异，故本文将历史数据划分为不同调度情景来训练模型，以期能得到不同调度情境下最优的预测结果。

灌区用水调度中调度员参考的流量调度指标众多。从表 3 可以看出，调度请求在不同调度场景中重要性程度均靠前，故从重要程度排名来看首先考虑调

度请求指标。参考调度员建议同时考虑到泾史杭灌区实际调度中各参数可获取频率，本文同时考虑较为重要、实际调度中可获取频率较高且随机性较大的降水指标。为此，依据这 2 个指标把所有数据划分为表 2 所示的 5 种场景。

表 2 调度情景划分

Table 2 The division of scheduling scenario

场景编号	场景描述	数据组数
1	所有特征数据集	2 192
2	无降水、无调度请求	408
3	无降水、有调度请求	200
4	有降水、无调度请求	1 152
5	有降水、有调度请求	432

1.3.2 SHAP 方法剔除冗余特征量

沙普利可加性特征解释方法（Shapley additive explanation，SHAP）可对模型特征变量重要性进行评估。灌区用水调度过程中，针对目标流量，管理人员事实上在不断权衡各影响要素或特征变量的重要程度，进而动态联合决策，这是一个典型的各特征变量之间的合作博弈过程。在合作博弈中，各特征变量的 Shapley 值兼具了效率和公平的原则，每个特征变量的 Shapley 值是所有可能特征变量组合对目标结果贡献的加权求和，其计算过程体现的是由附加特征变量带来影响后造成的预测值之间的差距。而 SHAP 值则是基于合作博弈中的最佳 Shapley 值，直接反映了特征变量的重要程度[11-12]，可由式（1）计算得到。每 1 个数据样本对应的预测值，都可得到 1 个 SHAP 值，其有正负，其绝对值反应的就是该特征的影响力大小。而特征变量的重要性赋分为所有 SHAP 值绝对值的平均[13-15]。

$$g(z')=\phi_0+\sum_{j=1}^{M}\phi_j z_j', \qquad (1)$$

式中：$z'\in\{0,1\}^M$是联合向量，$\{0,1\}^M$中的 0 与 1 代表了特征变量是否参与了该联合向量的计算；M 是变量维度；$\phi_j\in R$ 是特征 j 的 Shapley 值。

在得到特征变量的重要性赋分排名后，即可剔除得分较低的特征量，以此消除冗余特征量在模型预测时相互产生的影响，以提高模型预测精度。

1.3.3 随机森林模型的构建

随机森林模型（Random Forest，RF）是以树状网络的形式，形成灌区用水调度目标流量与各特征变量之间的非线性回归映射。该映射基于所有与灌区用水调度相关的特征变量作为样本集，在特定规则下形成决策树节点与目标变量之间的树状决策（即决策树），再由众多决策树形成决策森林网络。形成决策树的规则是，放回且随机在样本集中抽选出数目为 M 的特征变量，再从其中选取 m 个变量（$m<M$）作为决策树的节点，当对任意特征变量所选取的任意划分两侧的样本数据均方差和最小时，该决策树的节点将分化，由此递归形成最优决策树，且单棵树依据叶子节点的均值得到预测值。产生 N 棵决策树后，目标变量的最终预测结果为所有树预测值的平均值。在根据预测结果不断优化决策树数目后，得到最终的灌区用水调度目标流量与各特征变量之间的非线性回归映射[16-17]。

1.3.4 BP 神经网络模型的构建

BP（Back Propagation）神经网络，即基于误差反向传播算法的多层前馈神经网络（Multilayer Feedforward Network），同样能够对非线性问题进行很好的拟合描述。BP 神经网络由输入层、隐藏层和输出层 3 部分组成。将归一化后的灌区用水调度相关数据输入网络模型，网络中数据会被正向的从输入层经隐含层向输出层传递，输出层的各个神经网络会得到相应的响应。输出层各期望输出和实际输出值之间会形成误差，这个误差会被按照误差减小的方向重新从输出层向输入层方向反向传播，并在经隐含层时修正各个连接的权值和阈值，直至误差降低到可以接受的范围，使实际输出接近期望输出。即正向计算的输出，反向传播误差。在根据输出结果不断优化隐含层神经元数目后，可最终得到灌区用水调度目标流量与各特征变量之间的非线性回归映射。

1.3.5 基于 Neo4j 图形库的知识图谱搭建

知识图谱是一个具有向图结构的知识库，其概念自从由 Google 公司从 2012 年提出后，现已被应用于知识工程、自然语言处理、智能问答等诸多人工智能领域[17]，其构建过程主要包括知识获取、知识表示、知识储存以及知识可视化 4 个步骤。结构化、半结构化以及分结构化数据均可用来构建知识图谱[18-19]。本文的知识库，就是机器学习模型预测得到的调度流量以及其相应其他特征变量结构化的数据库。

Neo4j 是一个稳定且成熟的，具有较高性能的图形数据库，具有可保证数据完整性、可用性高、可扩展性强以及能够高速检索数据等特点，这些特点完美的契合了灌区调度知识图谱搭建的需求。

1.4 模型精度评价与验证

用训练期间未使用过的数据对模型进行测试可获得对性能更加客观的估计，故本文对 2 个模型的验证，均是随机从完整数据集中单独抽取 1/10 的数据进行对比验证，定量评价指标包括决定性系数（coefficient of determination，R^2）和平均绝对误差（mean absolute error，MAE）。

2 知识图谱查询体系构建过程

2.1 特征组合的选择

以 2015—2020 年共 2 192 组数据作为样本数据，

输入全部特征变量，利用 SHAP 方法得到每一个特征变量的重要性得分及排名。为统一定量对比分析，得分归一化后不同调度情景重要性赋分非零且排名前 20 的特征变量及排序如表 3。可见，不论是哪一种调度情景，温度和时间的重要性得分排名都较高，这与陆面温度和时间进程直接决定着作物生长密切相关，即调度会重点关注作物的生长情况。同时，凡是调度场景中有调度指令的下达，其排名和得分也较高，说明泾史杭灌区在日常调度中，凡是各调度单元提出的调度需求基本都会优先考虑，特别是六冲的需求电报，这与其位于瓦西干渠中下游有很大的关系。即调度员会优先考虑各调度单元提出的调度需求，渠道中下游是关注的重点。对蓄水量来说，无降水以及有用水需求时，调度员重点关注的是对灌域内旱涝情况更敏感的塘坝和小（二）型水库；而无用水需求时，除了塘坝蓄水量，同时还关注能对局部灌域起到调蓄作用的中型水库，特别是当有降水无用水需求时，中型水库排名显著上升。即从数据分析可看出，在没有降水和调度请求时，城市和生态供水补水水量重要性得分相较其他场景显著提升。降水量重要性相对较低，这与泾史杭灌区降水较为规律，日期在一定程度上即可代表降水变化，掩盖了其贡献程度有关。值得注意的是，降水的重要性得分及排名在有降水、无调度请求的调度场景中有所提升。

为了使模型预测效果更好，需要对输入模型的特征变量进行初步筛选，尽可能地剔除冗余的变量。最终选择数据样本中特征变量重要性赋分从高到低，归一化后累计和达到 90%的特征变量，作为最终输入预测模型的变量组合，各组合如表 4。

表 3 不同调度情景特征变量重要性得分排序

Table 3 Ranking of importance scores of feature variables in different scheduling scenarios

排名	无降水、无调度请求		无降水、有调度请求		有降水、无调度请求		有降水、有调度请求		所有特征数据集	
	特征变量	重要性得分	特征变量	重要性得分	特征变量	重要性得分	特征变量	重要性得分	特征变量	重要性得分
1	*T*1	0.403	*T*	0.411	*T*1	0.187	*YL*	0.470	*YL*	0.398
2	*T*2	0.122	*T*1	0.134	*T*2	0.171	*T*1	0.106	*T*1	0.161
3	*S*4	0.109	*S*4	0.096	*S*4	0.128	*S*3	0.095	*S*3	0.088
4	*T*	0.100	*S*3	0.091	*S*3	0.127	*T*	0.070	*T*	0.071
5	*S*2	0.063	*YL*	0.055	*T*	0.073	*T*2	0.043	*T*2	0.070
6	*S*3	0.062	*S*1	0.039	*T*3	0.061	*C*	0.034	*S*4	0.067
7	*C*	0.047	*T*2	0.033	*S*1	0.057	*T*3	0.027	*S*1	0.024
8	*S*1	0.045	*H*	0.032	*S*2	0.039	*S*1	0.022	*S*2	0.021
9	*H*	0.028	*T*3	0.032	*C*	0.021	*S*4	0.022	*C*	0.019
10	*T*3	0.018	*G*	0.026	*H*	0.018	*YW*	0.020	*H*	0.015
11	*G*	0.005	*C*	0.020	*R*49	0.008	*H*	0.014	*T*3	0.014
12	-	-	*YW*	0.018	*G*	0.007	*S*2	0.005	*YW*	0.009
13	-	-	*S*2	0.012	*R*58	0.007	*G*	0.005	*G*	0.004
14	-	-	-	-	*R*46	0.004	*R*29	0.004	*R*48	0.002
15	-	-	-	-	*R*18	0.004	*R*24	0.003	*R*49	0.002
16	-	-	-	-	*R*34	0.003	*R*4	0.002	*R*12	0.001
17	-	-	-	-	*R*48	0.003	*R*49	0.002	*R*36	0.001
18	-	-	-	-	*R*54	0.003	*R*15	0.002	*R*58	0.001
19	-	-	-	-	*R*39	0.002	*R*36	0.002	*R*38	0.001
20	-	-	-	-	*R*37	0.002	*R*54	0.002	*R*52	0.001

注 表中符号具体含义见表 1，下同。

表 4 不同调度情景选用的特征组合

Table 4 The combination of features selected for different scheduling scenarios

场景序号	特征变量组合	特征值重要性得分累加和
1	*YL*+*T*1+*S*3+*T*+*T*2+*S*4+*S*1+*S*2	0.901
2	*T*1+*T*2+*S*4+*T*+*S*2+*S*3+*C*	0.905
3	*T*+*T*1+*S*4+*S*3+*YL*+*S*1+*T*2+*T*3+*H*	0.923
4	*T*1+*T*2+*S*4+*S*3+*T*+*T*3+*S*1+*S*2+*C*+*H*+*R*49+*G*+*R*58	0.903
5	*YL*+*T*1+*S*3+*T*+*T*2+*C*+*T*3+*S*4+*S*1+*TW*	0.910

2.2 2 种模型不同调度场景验证对比及预测模型的确定

由于泾史杭灌区调度相关历史数据的数量有限，故本部分对比分析了不同调度情景下随机森林模型和 BP 神经网络预测灌区用水调度流量的准确度，最终择优选之，以期在训练样本有限的情况下得到预测值较为准确的调度场景数据库。图 2 为 2 种模型的构建示意图。

从图 3 纵向对比可知，在所有场景组合中，图 3（a）、图 3（b）的拟合线与 1∶1 线最为贴合，且 *MAE* 和 R^2 也最优。由此可知，通常情况下 BP 神经网络和随机森林模型在训练样本数量越多时，预测精度越高。故在实际应用时，虽然有不同调度场景可选择，但也应尽可能选择有较多训练样本的场景来匹配预测流量。同时，未来也应在扩充历史数据的前提下，不断优化已有模型精度。与此同时，当特征组合中特

征变量的重要性差异较小时（图 3（c）），即使训练样本数量足够大，剔除特征值冗余后，BP 神经网络最终的预测准确度也不会得到很大的改善，以至于出现了训练样本数量较大，但最终预测准确度不高的情况，如图 3（g）。

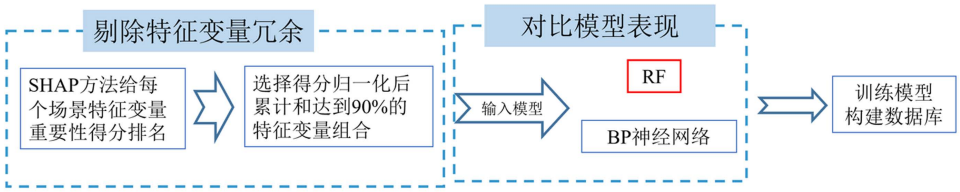
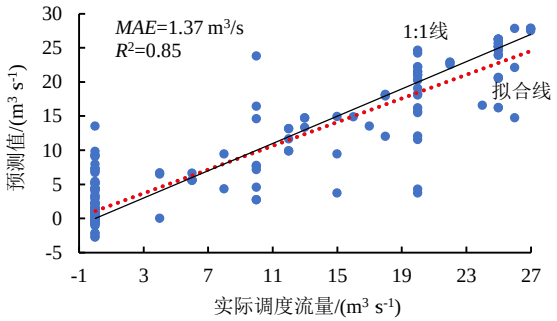
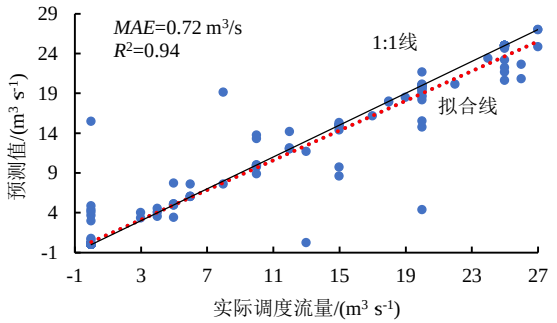


图 2 模型构建示意图

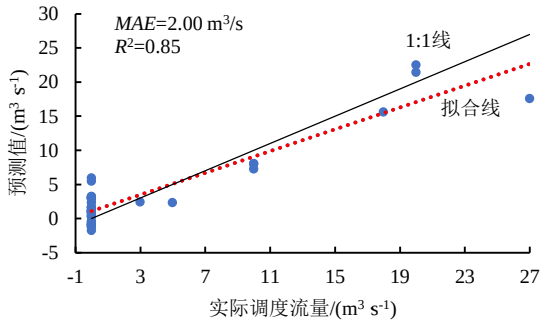
Fig.2 Model construction diagram



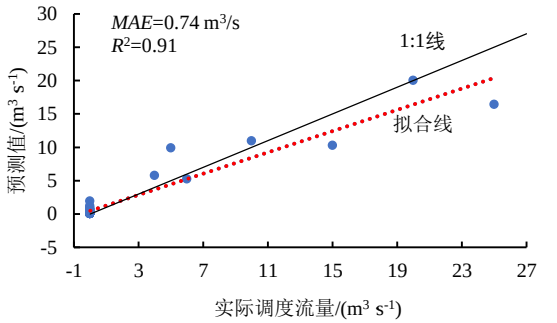
(a) 所有特征组合-BP 神经网络



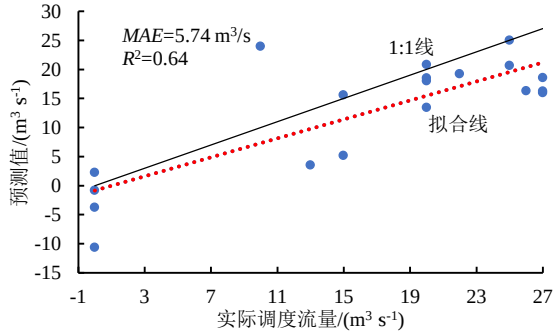
(b) 所有特征组合-随机森林模型



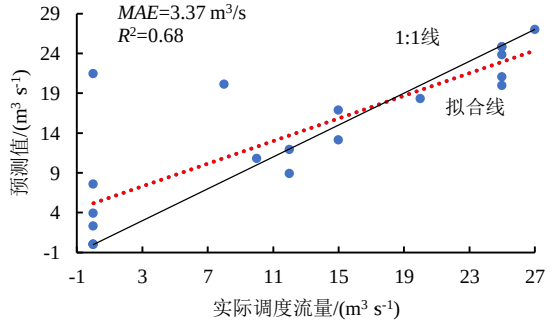
(c) 无降水、无调度请求-BP 神经网络



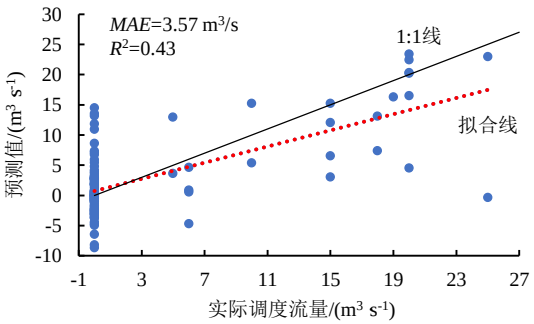
(d) 无降水、无调度请求-随机森林模型



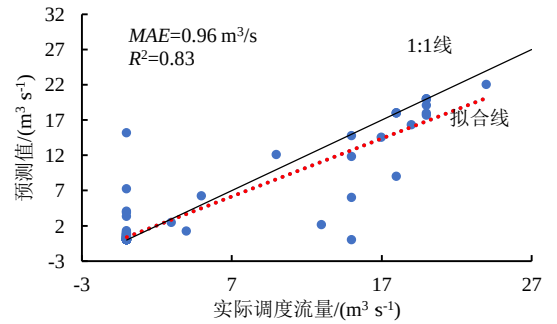
(e) 无降水、有调度请求-BP 神经网络



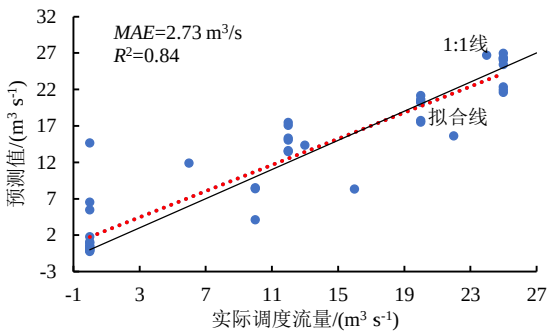
(f) 无降水、有调度请求-随机森林模型



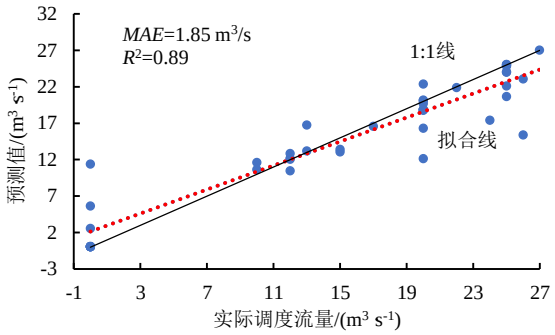
(g) 有降水、无调度请求- BP 神经网络



(h) 有降水、无调度请求-随机森林模型



(i) 有降水、有调度请求- BP 神经网络



(j) 有降水、有调度请求-随机森林模型

图 3 不同调度情景下 2 种模型预测结果对比

Fig.3 The comparison of prediction results of two models under different scheduling scenarios

横向对比可看出，随机森林模型所形成的灌区用水调度目标流量与各特征变量之间的非线性回归映射的预测结果明显优于 BP 神经网络，其平均绝对误差（*MAE*）和决定性系数（R^2）在不同调度情景下都更优，特别是在有降水、无调度请求这一情境下优势更明显。

整体而言，随机森林模型更适于当前历史数据体量下灌区用水调度流量预测，故本文选用随机森林模型构建灌区用水调度流量预测值数据库。

2.3 图谱查询体系的构建

2.3.1 数据库准备

为了使图谱中的场景库数据能够尽可能的覆盖现实调度场景，需得到合理且有一定密度的特征变量组合数据，即对现有的调度场景进行加密处理。具体的处理步骤为：首先分析泾史杭灌区 2015—2020 年已有历史数据各特征变量的出现规律，进而划分出各特征变量的数值梯度。参照表 4 的特征组合将不同特征变量的数值梯度进行排列组合，最终不同调度场景下的不同组合即为输入随机森林模型的预测集，以此可得到相应调度流量预测值，由此形成调度流量数据库样本，即得到了有一定密度的调度场景库。

特征变量具体划分梯度如表 5。其中日累计降水量的梯度划分参照降水量等级标准（GB/T 28592—2012）；六冲测流站以及瓦西进水闸调度请求的流量值划分由于历史数据重复性高且组合较少，故每一个历史调度请求值都单独划分为 1 组；温度按照每 5 ℃为 1 个梯度；罗管闸水位除了极少数关闸时期，其水位需常年维持，故 0 m 单独划分为 1 组。其他特征变量都按照历史出现值划分为 11 组。在考虑到降水和温度与时间具有较大相关性后，最终排列组合得到的数据组数为：所有特征数据集 8.34×10^8 组，无降水无调度请求 9.27×10^7 组，无降水有调度请求 4.02×10^9 组，有降水无调度请求 9.55×10^{11} 组，有降水有调度请求 2.21×10^{10} 组。将这些数据分别作为各调度情景对应训练好的随机森林模型预测数据的输入数据集，得到不同调度情境下各组合的调度流量预测值。即共有约 9.82×10^{11} 组数据构成泾史杭灌区调度流量预测知识库的样本数据。

表 5 特征变量梯度划分

Table 5 The gradient partitioning of characteristic variables

特征变量	划分梯度范围	划分梯度（/组）
T1/℃	-10 ~ -35	5
T2/℃	-10 ~ -40	5
T3/℃	-10 ~ -35	5
R/mm	-10 ~ 200	-
C/万 m^3	0 ~ 1 100	110
S1/万 m^3	0 ~ 860	86
S2/万 m^3	0 ~ 570	57
S3/万 m^3	0 ~ 880	88
S4/万 m^3	50 ~ 1 680	163
H/m	0，48.70 ~ 50.90	0.2
YL	-	-
YW	-	-
G	-	-

2.3.2 搭建图谱查询体系

本文以瓦西闸为突破点，构建适用于泾史杭灌区各个闸门用水调度流量预测的知识图谱模型，为此，图 4 首先给出了泾史杭灌区 28 个主要调控闸门的图谱结构，在此基础上，重点关注瓦西进水闸。将前文得到的调度流量预测值组合数据导入 Neo4j 图形库中，创建流量预测值（prediction）和与特征变量梯度组合（group）2 类节点，并按照对应组合建立二者之间的映射关系。其中，各梯度的具体数值作为节点 group 的属性值。进而，在代码层面将该映射关系的调度流量预测值组合与具有地理信息属性的瓦西进水闸节点关联起来，形成可实际查询的瓦西过闸流量调度知识图谱模型。若在具备实测数据的基础上，重复上述流程，即能构建起图 4 中其他闸门过闸流量调度的知识图谱模型。

实际应用上述构建的知识图谱模型时，依据实际调度情景在样本集中找到适宜的调度场景及表 5 中各特征值最接近的梯度值，可在图谱模型中检索得到调度目标对应的流量值。以 2020 年 8 月 20 日为例，当天有调度指令和降水，即可按照表 2 中场景 1（所

有特征数据集）和场景 5（无降水、有调度请求）进行检索，检索需要输入模型的各特征变量类型查表 4 可得，具体的数值根据实际所获数据参照表 5 找到最接近的梯度值即可。图 5 为从已构的知识图谱模型所有特征数据集（场景 5）调度情景中，检索 2020 年 8 月 20 日调度流量预测值的结果。在左上方代码输入框中输入包含各特征变量具体值的节点检索相关代码，即会检索到粉色节点，其上显示的数字是该节点的顺序编号。在点击粉色节点“扩展/折叠子关系”的功能按钮后即可显示出其关联的橘色节点和绿色节点，橘色节点上显示的值即为模型预测的调度流量值。其中橘色节点由于显示的是粉色节点变量组合对应的调度流量值，故其箭头指向粉色节点；而粉色节点是所有特征数据集中的一个节点，故箭头指向绿色节点，箭头表示的是节点间的包含关系，与检索顺序无关。

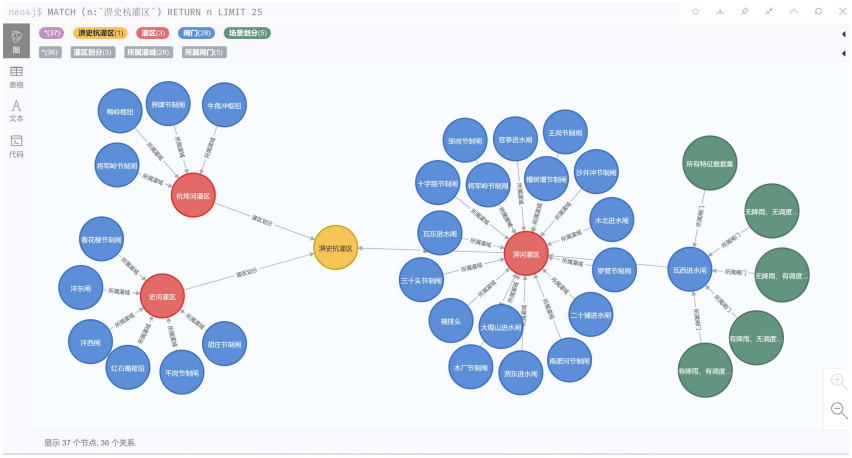


图 4 泖史杭灌区知识图谱主体框架

Fig.4 The main frame of knowledge graph in Pishihang Irrigation District

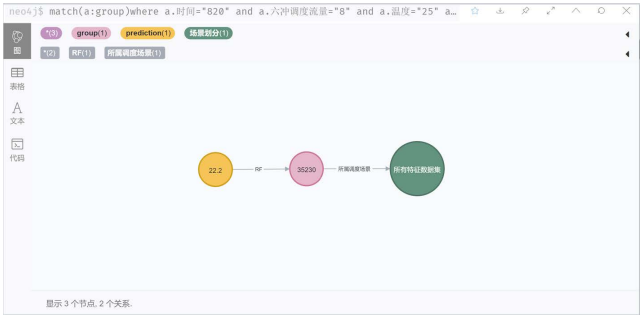


图 5 瓦西进水闸 2020 年 8 月 20 日调度流量预测值检索结果

Fig.5 The retrieval result of dispatch flow prediction value of Washi Inlet lock on 20200820

2.4 图谱查询体系的检验

为了验证上述构建的图谱查询体系精度，随机选择 2020 年 4 月 21 日、6 月 7 日、8 月 20 日、11 月 24 日 4 组数据验证。4 月 21 日无降水无调度请求，适用于场景 1 和场景 2；6 月 7 日无降水有调度请求，适用于场景 1 和场景 3；8 月 20 日有降水有调度请求，适用于场景 1 和场景 5；11 月 24 日有降水无调度请求，适用于场景 1 和场景 4。表 6 为验证结果的对比分析。从最后 2 列数据对比可以看出，从知识图谱模型中得到的流量与调度员下达调度指令实际调度的流量最大差异小于 3 m^3/s，在泖史杭灌区管理总局调度员认可的误差范围以内。值得关注的是，虽然该知识图谱模型中包含了 10^{11} 量级的节点数和更多的关系数，但调度流量的预测值检索过程仍具有实时性，故上述构建的知识图谱模型具有显著的实用性。由于在图谱构建前期的验证过程中，发现模型预测效果与训练集数据量有着极大关系，且部分场景模型预测效果并不佳。故在实际检索时，推荐用训练集数据量较大的场景进行流量预测检索。推荐选用场景 1 数据节点进行检索查询。

表 6 知识图谱查验结果对比

Table 6 The comparison of knowledge graph inspection results

时间	场景编号	检索流量	实际流量
0421	1	0.4	0
	2	0.1	0
0607	1	20.2	20
	3	19.46	20
0820	1	22.2	20
	5	18.1	20
1124	1	0.1	0
	4	1.64	0

3 讨 论

就调度流量值预测方法来看，针对影响灌区用水调度流量的变量类型具有属性迥异、且依赖于历史的基本特征，本文选取了人工智能技术中的相关模型开展了分析。人工智能技术中的随机森林模型和 BP 人工神经网络都具有较强的拟合回归高维非光滑、甚至非连续函数的能力，但该能力依赖于特定的物理问题

和应用场景。故本文在历史记录数据量有限的情况下对比分析了随机森林模型和 BP 神经网络在表征灌区用水调度流量和影响该流量的特征变量之间非线性映射方面的精度。结果为随机森林模型表现更优，这与 Ben 等[13]的结果相符。但更多的人工智能模型仍待用于灌区用水调度场景的精度对比分析，以筛选出适用场景更加复杂的最佳非线性映射表征算法。

从应用层面来说，本文所构建的图谱查询体系其本质上可以看作是一个简单的灌溉管理专家系统。灌溉管理专家系统的概念 Srinivasan 等[20]在 1991 年就提出过，研究者希望通过其来帮助调度员实现更加科学的调度。近年相关研究已取得进展：Wang 等[21]介绍了基于 Web 的调度决策支持系统，其能够实现动态配水决策和灌溉信息管理等功能，并在湖北省漳河灌区进行了实际的应用。这种系统搭建方法已成为近年来数字灌区建设的主流，与本文构建的查询体系相比，它的功能更齐全，且能实现实时的监测和计算。但由于其在实际应用时必需依赖网站界面的搭建和应用系统的集成开发，故在同样能实现实时提供适宜调度流量这一功能的情况下，本文所提供的方法极大地减少了系统搭建成本。

从创新性来说，知识图谱常见用于构建行业知识库[22]或进行某一领域的行业发展现状及趋势研究[23]。而本文所构建的图谱查询体系，则是利用了 Neo4j 图形库在构建知识图谱时可进行数据存储、实时检索以及图谱展示的功能，搭建起了一个连接模型算法与实际应用的桥梁，可极大的解决算法模型在学术层面不断优化，但应用转化率低的问题，即为之后算法模型的应用验证提供了一个新的实现思路。

应注意的是，当前灌区用水调度历史实测数据的数据量还无法让应用更为广泛的深度学习模型充分发挥其性能，故在进一步丰富实测数据的同时，未来拟结合物理模型计算所得数据以补充用水调度历史数据的数据集。同时，由于历史调度的科学性会很大程度上影响本文所构建方法的调度结果，故在有更加详实的调度记录情况下，若能在模型训练前对样本数据进行评价和筛选，提高训练集的科学性，可得到更好地调度效果。探索更加合理、精准的各特征变量梯度的划分方法也是后期需要完善的地方。调度最终是为作物生长服务的，如果调度员的视角最终能延伸到田间，将作物的生长情况纳为参考指标，在现有条件下加入遥感等监测数据，将会实现更加精准的灌溉调度。或将作物生长模型与调度模型相结合也是未来可能的一种发展方向。最后，在充分考虑极端突发事件发生的前提下，系统分析及验证机器学习模型在表征灌区用水调度流量和影响该流量的特征变量之间非线性映射的精度，也是未来的探索方向。

本文构建的灌区用水调度流量预测值知识图谱模型具有一定准确性和实用性，但仍需进一步探索和完善，以充分适应灌区常规和非常规应急用水调度问题。

4 结 论

本文在现有调度历史数据的基础上利用机器学习模型和 SHAP 方法丰富和加密了调度场景，得到了能实现基本覆盖现实调度场景的调度场景库。基于该场景数据库，利用 Neo4j 图形库构建出灌区用水调度流量预测值知识图谱模型，该模型能在一定精度下满足实际调度需求，可为灌区用水调度管理提供便捷的分析参考工具，为搭建实时、便捷、使用成本低的灌溉管理专家系统提供了新的实现思路。

（作者声明本文无实际或潜在利益冲突）

参考文献：

[1] RASCHKA S. Python machine learning[M]. Birmingham: Packt Publishing - ebooks Account, 2015.

[2] WOLPERT D H. The lack of a priori distinctions between learning algorithms[J]. Neural Computation, 1996, 8(7): 1 341-1 390.

[3] WOLPERT D H, MACREADY W G. No free lunch theorems for optimization[J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67-82.

[4] 徐建新, 张亮, 邱林, 等. 用改进的 BP 算法预报灌区地下水位[J]. 华北水利水电学院学报, 2003, 24(1): 1-4.
XU Jianxin, ZHANG Liang, QIU Lin, et al. Forecast groundwater level in irrigation district based on improved BP algorithm[J]. Journal of North China Institute of Water Conservancy and Hydroelectric Power, 2003, 24(1): 1-4.

[5] 刘扬, 任振辉, 谢景新, 等. 遗传算法双层次优化模型在精细灌溉决策中的应用[J]. 农业机械学报, 2007, 38(6): 125-128, 133.
LIU Yang, REN Zhenhui, XIE Jingxin, et al. Application and research on genetic algorithm with two layers optimize pattern in the precision water-saving irrigation decision[J]. Transactions of the Chinese Society for Agricultural Machinery, 2007, 38(6): 125-128, 133.

[6] 刘照, 华庆伟, 张成才, 等. 基于 RS、GIS 及智能算法的渠系优化配水[J]. 西北农林科技大学学报(自然科学版), 2017, 45(4): 213-222, 229.
LIU Zhao, HUA Qingwei, ZHANG Chengcai, et al. Optimal irrigation water distribution based on RS, GIS and intelligent algorithms[J]. Journal of Northwest A & F University (Natural Science Edition), 2017, 45(4): 213-222, 229.

[7] 陈述, 邵东国, 李浩鑫, 等. 基于粒子群人工蜂群算法的灌区渠-塘-田优化调配耦合模型[J]. 农业工程学报, 2014, 30(20): 90-97.
CHEN Shu, SHAO Dongguo, LI Haoxin, et al. Coupled allocation model for optimizing water in canal-pond-field based on artificial bee colony and particle swarm algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(20): 90-97.

[8] 卢麾, 田富强, 胡和平, 等. 基于遗传算法和 GIS 技术的灌溉决策支持系统[J]. 水利水电技术, 2002, 33(7): 27-30, 66.
LU Hui, TIAN Fuqiang, HU Heping, et al. Irrigation decision support system based on genetic algorithm and geographic information system[J]. Water Resources and Hydropower Engineering, 2002, 33(7): 27-30, 66.

[9] 赵丽华, 徐立中. 基于明渠非恒定流仿真模型的灌区水情信息耦合

研究[J]. 水利学报, 2012, 43(5): 537-544.
ZHAO Lihua, XU Lizhong. Hydrological information coupling for irrigation district based on numerical simulation of open channel unsteady flows[J]. Journal of Hydraulic Engineering, 2012, 43(5): 537-544.
[10] 朱红. 浭史杭灌区水资源配置与调度实践[J]. 中国水利, 2019(3): 16-18.
ZHU Hong. Water allocation and regulation in Pishihang Irrigation District[J]. China Water Resources, 2019(3): 16-18.
[11] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions[C]. Advances in Neural Information Processing Systems, 2017: 4 765-4 774.
[12] WEBBER J. A programmatic introduction to Neo4j[C]//Proceedings of the 3rd annual conference on systems, programming, and applications: software for humanity. New York: ACM, 2012: 217-218.
[13] BEN JABEUR S, MEFTEH-WALI S, VIVIANI J L. Forecasting gold price with the XGBoost algorithm and SHAP interaction values[J]. Annals of Operations Research, 2021: 1-21.
[14] LUNDBERG S M, ERION G G, LEE S I. Consistent individualized feature attribution for tree ensembles[J]. arXiv preprint arXiv: 1802.03888, 2018.
[15] MOLNAR C. Interpretable machine learning[M]. Lulu Press, 2019.
[16] GORDON A D, BREIMAN L, FRIEDMAN J H, et al. Classification and regression trees[J]. Biometrics, 1984, 40(3): 874.
[17] BERK Richard A. Classification and Regression Trees (CART)[M]. New York: Springer, 2016.
[18] 漆桂林, 高桓, 吴天星. 知识图谱研究进展[J]. 情报工程, 2017, 3(1): 4-25.
QI Guilin, GAO Huan, WU Tianxing. The research advances of knowledge graph[J]. Technology Intelligence Engineering, 2017, 3(1): 4-25.
[19] 王昊奋, 漆桂林, 陈华钧. 《知识图谱：方法、实践与应用》[J]. 自动化博览, 2020, 37(1): 7.
[20] SRINIVASAN R, ENGEL B A, PAUDYAL G N. Expert system for irrigation management (ESIM)[J]. Agricultural Systems, 1991, 36(3): 297-314.
[21] WANG W C, CUI Y L, LUO Y F, et al. Web-based decision support system for canal irrigation management[J]. Computers and Electronics in Agriculture, 2019, 161: 312-321.
[22] 王得强, 吴军, 关立文. 结合知识图谱的行业知识库构建方法研究[J]. 制造技术与机床, 2022(8): 74-80.
WANG Deqiang, WU Jun, GUAN Liwen. Industry knowledge base construction based on knowledge graph[J]. Manufacturing Technology & Machine Tool, 2022(8): 74-80.
[23] 陈江涛, 吕建秋. 基于知识图谱的运筹学发展现状及趋势研究[J]. 运筹与管理, 2019, 28(1): 194-199.
CHEN Jiangtao, LYU Jianqiu. Research on the present situation and tendency of operational research based on knowledge map[J]. Operations Research and Management Science, 2019, 28(1): 194-199.

Construction of Knowledge Graph Model for Irrigation Water Scheduling

SU Nan[1], ZHANG Shaohui[1,2*], BAI Meijian[1,2], ZHANG Baozhong[1,2]

(1. China Institute of Water Resources and Hydropower Research, Beijing 100038, China;

2. State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, Beijing 100038, China)

Abstract: 【Background and Objective】The adoption of artificial intelligence technology, particularly through adaptive learning, is reshaping the landscape of intelligent construction in the field of irrigation. This innovative perspective is divergent from traditional research paradigms, as numerous intelligent algorithms and technologies have been employed in the realm of irrigation water dispatching decisions. However, existing results often lack user-friendliness and ease of operation, preventing them from practical implementation. In light of this, this paper undertakes a comparative analysis of the Random Forest model and the BP Neural Network model in artificial intelligence technology.【Method】Building on this insight, the study divides historical data, based on measured values, into various gradients. These gradients encapsulate water irrigation scheduling variables, resulting in 9.82×10^{11} water dispatching scenarios. A pre-trained Random Forest model is employed to predict the dispatching flow corresponding to each scenario, ultimately constructing a fundamental database mapping the relationship between irrigation area water dispatching scenarios and dispatching flow values. This process enriches and encodes the scheduling scenario database through the utilization of machine learning models and existing scheduling history data. Leveraging this scenario database, and Neo4j technology, a knowledge map-based water dispatching flow prediction model is established for the Pishihang irrigation district.【Result and Conclusion】The key advantage of this model is its user-friendly approach. Irrigation area water scheduling managers can effortlessly determine the approximate gradient value of each scheduling variable within the knowledge graph model for a specific scheduling scenario. By doing so, they can readily access the corresponding scheduling flow prediction value through a simple query. Application results demonstrate that the knowledge map-based scheduling flow prediction model boasts a remarkable accuracy within the cognitive range of Pishihang irrigation district water management, facilitating real-time scheduling flow inquiries. This research represents a significant step toward more user-friendly and effective irrigation water management.

Key words: Pishihang Irrigation District; water dispatching; machine learning; SHAP method; knowledge graph

责任编辑：赵宇龙