严酷环境下岩土工程灾害防治技术及工程应用专栏

文章编号:1005-0930(2024)06-1664-014 中图分类号:P468.0+25 文献标识码:A
doi:10.16058/j.issn.1005-0930.2024.06.010

基于机器学习算法构建新疆积雪覆盖率预测模型

邓文彬, 侯雪晴
(新疆大学建筑工程学院,新疆 乌鲁木齐 830046)

摘要:积雪作为宝贵的淡水资源,其覆盖率的变动对农牧业经济的发展具有深远影响.当前对积雪覆盖率的预测研究较少,为提升积雪覆盖率预测的准确性,基于机器学习算法,构建支持向量回归(SVR)、粒子群(PSO)优化SVR、随机森林(RF)、XGBoost及优化后的XGBoost预测模型对新疆积雪覆盖率进行预测研究,并对模型预测精度进行对比分析.研究结果表明:RF和优化后的XGBoost模型的R^2均大于0.9;传统SVR模型的R^2均小于0.8,而PSO算法优化后的SVR模型的R^2均大于0.8,部分大于0.9;XGBoost模型的R^2均低于0.4.说明RF、优化后的XGBoost及PSO-SVR模型在积雪覆盖率预测研究中呈现出较高精度,XGBoost模型的预测结果最差,且利用不同算法对传统模型进行优化在研究中十分必要.

关键词:积雪覆盖率;SVR;粒子群优化算法;RF;XGBoost;参数寻优

积雪作为冰冻圈的关键组成部分,其累积与消融的动态过程深刻影响着能量循环和水资源的合理分配,同时在调节地表辐射平衡方面发挥着不可忽视的作用[1].作为丰富的淡水资源,积雪对于农业灌溉和畜牧业发展至关重要,为地区径流提供了稳定的补给[2].新疆地处我国西北地区,不仅是我国陆地上积雪覆盖面积最大的区域[3],也是中亚地区典型的干旱与半干旱气候代表区.山地冰川和季节性积雪提供了宝贵的淡水资源,也带来了雪灾和融雪型洪水等潜在风险,对牧区生态环境构成了严重威胁.积雪覆盖率的变化将直接作用于地区径流的形成和水资源循环的维持,进而对区域气候产生显著影响.因此,对新疆积雪覆盖率进行持续监测和精准预测,对保障水资源安全、防范自然灾害以及促进区域可持续发展具有重要意义.积雪覆盖率的变化受多种复杂因素共同作用,对其进行精确预测是一项极具挑战性的任务.

收稿日期:2024-07-01;**修订日期:**2024-10-15
基金项目:新疆维吾尔自治区自然科学基金项目(2022D01C55)
作者简介:邓文彬(1977—),男,博士,教授.E-mail:125864110@qq.com
通信作者:侯雪晴(1998—),女,硕士.E-mail:houxueqing1210@163.com

近年来,国内外学者开展了对不同地区积雪预测的研究.时兴合等[4]对前期累积雪量与5~9月份74个环流因子之间的相关性进行了统计,得出18个具有显著意义的检验因子,并利用最优子集进行仿真,得到最优预测公式,其准确率达85.1%,但受个别因素影响,某些年的预报结果并不理想;成菲等[5]采用BCC—CSM1.1m模型,利用1984~2019年的历史回算资料,对欧亚地区1月份和4月份的降雪过程进行了较为客观的评价,并对模型预测的误差成因进行了分析;郝靖宇等[6]利用气象数据构建支持向量回归(SVR)模型,预测了天山山区的积雪覆盖率,结果表明积雪覆盖率受多种气象因素综合影响,而单一因素的影响较为有限.目前对新疆积雪覆盖率的预测研究较少,且存在预测数据不全面、算法不够丰富、预测精度有待提升等问题.

随着科学技术不断发展,机器学习发展迅速、种类繁多,其应用范围十分广泛,但各模型均存在各自的优缺点.如:支持向量机(SVM)具备严谨的数学逻辑,在小型数据样本的预测方面有着较高精度,其强大的学习和泛化能力使其在回归问题的发展中形成的支持向量回归(SVR)预测模型在气象、建筑等领域获得了广泛应用[7-9];传统SVR模型存在参数选取时间过长且必须依赖人工经验,难以将损失降到较低水准;粒子群算法(PSO)可有效减少搜索时间,以较快速度得到SVR模型的最优参数,从而提高预测精度.近年来,学者们将PSO算法与机器学习相结合,在不同领域取得了丰富的成果[10-11];随机森林(RF)模型采用bagging算法,通过有放回的抽样构建多棵决策树,以多棵树的集成得到最终结果,因其具有较高的准确性也被广泛应用于生态研究、材料和采矿领域[12-14],但在小型或高维度数据样本的预测中容易产生过拟合现象;极限梯度提升回归(XGBoost)是通过多个弱学习器来构建一个强大的模型,预测精度较高,被广泛应用于电力和灾害研究等领域[15-16],但其对参数较为敏感,参数不适宜容易导致结果过拟合或欠拟合.基于上述研究,以气象因子作为特征参数建立了不同的积雪覆盖率预测模型,为机器学习在不同领域的发展及对新疆地区积雪资源的预测方面提供一定的理论支持.该研究的主要贡献及创新点包括:①参考相关文献选取15个气象因素作为特征参数,以新疆2000年3月至2022年6月的积雪覆盖率为样本数据,建立了不同的积雪覆盖率预测模型,反映出机器学习算法在新疆积雪覆盖率预测方面的可行性及必要性.②基于机器学习的发展现状,结合各模型的优缺点,分别建立了SVR、RF和XGBoost预测模型,并对SVR和XGBoost模型进行参数优化,对比分析各模型的预测精度,进一步展示出不同机器学习算法在积雪预测领域的强大作用,也反映出对传统算法进行优化的必要性.

1 基本理论与研究方法

1.1 支持向量机回归(SVR)

支持向量机作为一种监督学习模型[17],由Corinna Cortes等[18]于1995年提出,旨在应用于数据分类、回归分析及孤立点监测等领域.该算法的核心目标在于在n维空间中寻找一个能够清晰划分数据点的超平面[19].在回归问题中被进一步发展为SVR[20].其目标是使回归超平面尽可能平滑,超平面越平滑,表示SVR模型的泛化能力越强[21].

SVR模型具有稀疏性,当样本点接近于回归模型时,损失可忽略.此时损失函数为不敏感损失函数$L(z)=\max(0,|z|-\varepsilon)$.该损失函数与SVM中的铰链损失函数相似[22],原点

附近值固定为 0.SVR 通过添加松弛变量来表示样本偏离“管道”程度,而“管道”是指为确保回归模型与预测值的偏差在特定范围内,对间隔加以限制.据此,SVR 优化问题可表达为

$$\max_{\omega,b}\frac{1}{2}\|\omega^2\|+C\sum_{i=1}^{N}(\xi_i+\xi_i^*) \tag{1}$$
$$\text{s.t.}\quad y_i-f(x)\leqslant\varepsilon+\xi_i$$
$$f(x)-y_i\leqslant\varepsilon+\xi_i^*$$
$$\xi\geqslant 0,\xi^*\geqslant 0$$

式中:C 为惩罚因子;ξ_i 和 ξ_i^* 均为松弛变量.

类似于软间隔 SVM,引入拉格朗日乘子,可得到拉格朗日函数和对偶问题

$$L(\omega,b,\xi,\xi^*,\alpha,\alpha^*,\mu,\mu^*)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}(\xi_i+\xi_i^*)-\sum_{i=1}^{N}\mu_i\xi_i-\sum_{i=1}^{N}\mu_i^*\xi_i^*+$$
$$\sum_{i=1}^{N}\alpha_i[f(x_i)i-y_i-\varepsilon-\xi_i]+\sum_{i=1}^{N}\alpha_i^*[f(x_i)i-y_i-\varepsilon-\xi_i^*] \tag{2}$$
$$\max_{\alpha,\alpha^*}\sum_{i=1}^{N}[y_i(\alpha_i^*-\alpha_i)-\varepsilon(\alpha_i^*+\alpha_i)]-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[(\alpha_i^*-\alpha_i)(x_i)^T(x_j)(\alpha_j^*-\alpha_j)]$$
$$\text{s.t.}\quad\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)=0,\ 0\leqslant\alpha_i,\alpha_i^*\leqslant C$$

式中:α、α^*、μ、μ^*均为拉格朗日乘子.

其中,对偶问题符合如下 KKT 条件[23]

$$\begin{cases}\alpha_i\alpha_i^*=0,\xi_i\xi_i^*=0\\(C-\alpha_i)\xi_i=0,(C-\alpha_i^*)\xi_i^*=0\\\alpha_i[f(x)-y_i-\varepsilon-\xi_i]=0\\\alpha_i^*[y_i-f(x)-\varepsilon-\xi_i^*]=0\end{cases} \tag{3}$$

对该对偶问题进行求解,得到 SVR 的形式为

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)x_i^Tx+b \tag{4}$$

支持向量回归的核函数有多项式函数、径向基核函数(RBF)、线性核函数以及 sigmoid 核函数[24].径向基函数可有效地处理高维数据,有着良好的非线性表达能力,同时其具有较强的泛化能力及鲁棒性[25],因此在建立 SVR 模型时选择径向基核函数.惩罚因子 C 值过大或过小会导致预测结果过拟合或欠拟合,而核函数参数 g 值的选择会对算法的预测精度产生直接影响.采取 PSO 算法进行 SVR 模型优化,以提高预测精度.

1.2 粒子群优化算法

粒子群算法(Particle Swarm Optimization,PSO)是在 1995 年由美国 James Kennedy 和 Russell Eberhart 共同提出的,并运用了 Hepper 的生物群体行为模式[26],即对鸟群中的个体找寻附近食物规律的研究[27].粒子群算法的基本思想是模拟鸟类通过自己的经验和群体之间的沟通来调节飞行路线,以找到食物的行为.该算法是一种基于粒子之间的协同作用,在多维空间中寻找最优解的一种方法[28].在 PSO 算法中,粒子是优化问题的关键,其用特征位置和速度来描述,首先初始化所有粒子的位置和速度,然后根据粒子的适应度函

数值,通过迭代搜索来获取最优解[29].

粒子群算法的原理:假设有 N 个粒子在一个 D 维空间进行搜索,所有的粒子存在一个由适应度函数确定的适应值,用于判断当前位置的好坏,各粒子均能记忆搜索过的最佳位置,且速度可依据自身及种群飞行经验进行动态调整.

粒子 i 的位置:$x_i=(x_{i1},x_{i2},\cdots,x_{iD})$,将 X_i 带入适应度函数中求得适应值;

粒子 i 的速度:$v_i=(v_{i1},v_{i2},\cdots,v_{iD})$;

粒子 i 个体经过的最佳位置:$pbest_i=(p_{i1},p_{i2},\cdots,p_{iD})$;

种群经过的最佳位置:$gbest=(g_1,g_2,\cdots,g_D)$.

在第 $d(1<d<D)$ 维中,位置变化在$[x_{\min,d},x_{\max,d}]$范围内,速度变化在$[v_{\min,d},v_{\max,d}]$范围内.

则有粒子 i 在 d 维速度的更新公式为

$$\boldsymbol{v}_{id}^{k}=\omega\boldsymbol{v}_{id}^{k-1}+c_1r_1(pbest_{id}-\boldsymbol{x}_{id}^{k-1})+c_2r_2(gbest_d-\boldsymbol{x}_{id}^{k-1}) \tag{5}$$

粒子 i 在 d 维位置的更新公式为

$$\boldsymbol{x}_{id}^{k}=x_{id}^{k-1}+v_{id}^{k-1} \tag{6}$$

式(5)、式(6)中:$\boldsymbol{v}_{id}^{k}$为粒子 i 在第 k 次迭代后飞行速度矢量的第 d 维分量;$\boldsymbol{x}_{id}^{k}$为粒子 i 在第 k 次迭代后位置矢量的第 d 维分量;c_1、c_2 为加速度常数,即调节学习的最大步长;r_1、r_2 为[0,1]范围内的随机数;ω 为惯性权重因子,是一个非负数.

从公式(5)可以看出,每个粒子速度的更新受 3 方面影响:①粒子的初始速度,作用于平衡全局和局部搜索;②粒子本身的记忆,使其具有全局搜索的能力;③群体信息,使得粒子间能够进行信息共享[30].三方面共同作用,使各粒子可以根据自身飞行经验及粒子间的信息共享不断调整更新自己的位置,以此来获取最优解.

1.3　随机森林回归(RF)

随机森林算法(Random Forest,RF)是一种创新的机器学习技术,基于分类树构建,由 Leo Breiman 提出[31].该方法以传统决策树为基础,在延续其优势的基础上,融入了集成学习理念,有效应对复杂且高度相关的特征分析[32].在处理包含缺失值的数据时,随机森林算法展现出卓越的鲁棒性[33],同时有效规避了过度拟合的风险.

RF 模型采用 Bagging 算法,即通过随机采样的方式获取各决策树的训练样本和构建决策树的特征,通过对多个决策树投票得到最终的预测结果.随机森林回归模型如图 1 所示.

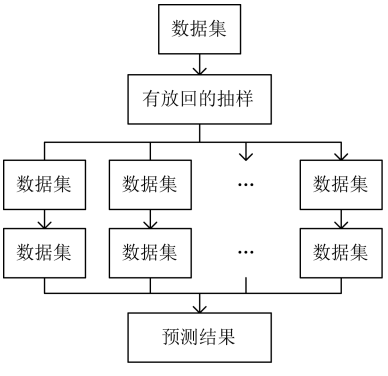


图 1　随机森林回归模型示意

Fig.1　Random forest regression model

1.4　极限梯度提升回归(XGBoost)

极限梯度提升(XGBoost)算法是另一种集成学习算法[34],其核心思想在于运用梯度提升技术,逐步训练多个弱学习器,在每一轮迭代中修正上一轮模型的残差,采用加权学习和梯度下降的方式优化各弱学习器,从而构建一个高效的集成学习模型[35].对损失函数采用二阶泰勒级数展开,同时加入正则项,可有效避免模型的过拟合.

XGBoost 是由 k 个基模型组成的一个加法运算式,其预测结果通式为

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{k} f_k(x),\ f_k \in F \tag{7}$$

式中:$\hat{y}_i$ 为预测结果;$\phi(x_i)$ 为样本 x_i 的预测分数;k 为树的总棵数;f_k 为第 k 棵决策树;F 为决策树对应的函数空间.

定义损失函数为

$$Obj(x) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{k} \Omega(f_i),\ f_i \in f \tag{8}$$

式中:$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ 为样本 x_i 的训练误差;$\Omega(f_i)$ 为第 j 棵树的正则项.

二阶泰勒级数展开的优化后的目标函数和 *Gain* 函数分别为

$$Obj(x) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{9}$$

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda}\right) \tag{10}$$

式中:*Gain* 为树分裂后目标函数的损失值;G_j 为损失函数的一阶导数之和;H_j 为损失函数的二阶导数之和;γT 为惩罚项;G_L、G_R 分别为左、右损失函数一阶导数之和;H_L、H_R 为分别为左、右损失函数二阶导数之和;λ 为惩罚系数.

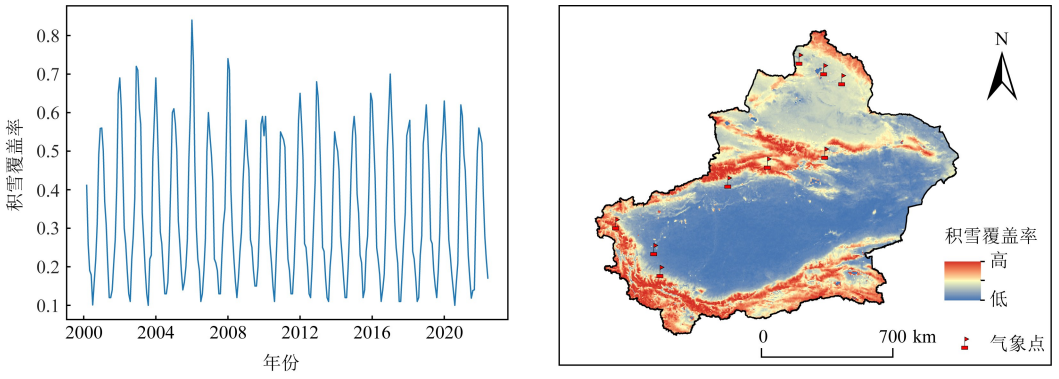
2 数据及来源

2.1 目标数据

积雪覆盖率的变化与气象因素密切相关,是多种因素综合影响产生的变化[36].利用美国国家冰雪数据中心(National Snow and Ice Data Cente,NSIDC)提供的 Terra 卫星监测的月合成积雪覆盖产品,反映当月最大的积雪覆盖范围,空间分辨率为 500m,数据格式为 hdf,覆盖研究区的有 6 幅,轨道号分别为 h23v04、h23v05、h24v04、h24v05、h25v04 和 h25v05,利用 MRT 软件进行数据的拼接、投影与格式转换,最终形成逐月积雪覆盖栅格数据.利用 matlab 软件提取逐月积雪覆盖率数据,以表格形式存储,去除 4 个月份的数据缺失,共有样本数据 264 个(数据展示见图 2(a)).

新疆地处我国西北边陲,山脉与盆地相间排列,有着“三山夹两盆”的独特地形,北部的阿尔泰山、南部的昆仑山、中部的天山是永久积雪分布的主要区域.鉴于单一气象站点数据在构建预测模型时可能存在的非确定性以及由此引发的偶然性风险,为确保预测结果的准确性及稳定性,选取分布于这三大山脉附近的 9 个关键气象站点的气象数据作为特征值纳入考量范畴,包括阿勒泰、哈巴河、富蕴、和静、达坂城、拜城、莎车、皮山以及乌恰地区(站点分布与积雪覆盖情况见图 2(b)).

结合机器学习算法的发展,分别建立各站点及 9 个站点均值的 SVR 模型、RF 模型、XGBoost 模型、优化 XGBoost 模型和 PSO-SVR 模型对新疆积雪覆盖率开展预测研究,并对各模型进行精度分析,避免了单一站点带来的偶然性,使得对新疆地区的积雪覆盖率预测更加科学、精准.



（a）2000年3月~2022年6月新疆积雪覆盖率的变化　　（b）气象点分布

图 2　各年积雪覆盖率与气象点分布

Fig.2　Distribution of snow cover and weather points by year

2.2　特征参数的选择

积雪的积累与消融受多种气象因素影响,如气温低、降水充足、日照时长较短,积雪消融速度缓慢,从而导致积雪覆盖面积增加;太阳辐射通过直接影响热力与水文过程而间接影响积雪的消融[37];大风伴随着气温降低,使得积雪难以消融[38].综合考虑,选取以下15个气象因素作为特征因子:距地表2m的气温、地表温度、露点温度、海平面气压、地面气压、潜在蒸发量、净日照强度、日照时数、蒸发量、总日照强度、直接辐射、相对湿度、风速、降雪量以及积雪深度.

本文所引用的气象数据均源自美国国家海洋和大气管理局(NOAA)气候预测中心(CPC)(NOAA-Climate Prediction Center,https://www.cpc.ncep.noaa.gov/).

为了降低弱相关因子对预测模型精度的不良影响,首先计算各站点特征因子与预测目标间的皮尔逊相关系数 P(图 3).

将皮尔逊相关系数的绝对值按大小依次分为极强相关、强相关、中等相关、弱相关和

特征因子	阿勒泰	哈巴河	富蕴	和静	达坂城	拜城	莎车	皮山	乌恰
距地表2m的平均气温/℃	−0.94	−0.95	−0.96	−0.96	−0.96	−0.96	−0.96	−0.96	−0.96
地表温度/℃	−0.93	−0.95	−0.95	−0.95	−0.95	−0.96	−0.95	−0.96	−0.95
露点温度/℃	−0.93	−0.94	−0.95	−0.96	−0.95	−0.95	−0.9	−0.9	−0.96
海平面气压/hPa	0.9	0.91	0.88	0.89	0.89	0.89	0.9	0.9	0.8
地面气压/hPa	0.87	0.88	0.85	−0.51	−0.57	0.38	0.83	0.68	−0.83
潜在蒸发量/mm	0.87	0.88	0.88	0.9	0.91	0.89	0.89	0.9	0.9
净日照强度(net)/(J/m^2/d)	−0.85	−0.86	−0.87	−0.89	−0.89	−0.87	−0.84	−0.84	−0.89
日照时数(峰值)/h	−0.83	−0.83	−0.83	−0.79	−0.82	−0.79	−0.83	−0.82	−0.81
蒸发量/mm	0.83	0.75	0.65	0.87	0.85	0.83	0.88	0.54	0.85
总日照强度/(J/m^2/d)	−0.82	−0.82	−0.82	−0.76	−0.8	−0.77	−0.83	−0.81	−0.84
直接辐射/(J/m^2)	−0.81	−0.81	−0.82	−0.76	−0.8	−0.77	−0.83	−0.81	−0.84
相对湿度/%	0.79	0.81	0.84	0.21	0.27	0.43	0.59	0.56	0.3
风速10m/(m/s)	−0.69	0.57	−0.56	0.66	−0.79	−0.59	−0.79	−0.84	−0.42
降雪量/mm	0.65	0.65	0.6	0.2	0.15	0.47	0.55	0.58	0.12
积雪深度/mm	0.52	0.46	0.59	0.6	0.65	0.61	0.31	0.49	0.49
$\|1\|>P\geq\|0.8\|$	极强相关	$\|0.8\|>P\geq\|0.6\|$	强相关	$\|0.6\|>P\geq\|0.4\|$	中等相关	$\|0.4\|>P\geq\|0.2\|$	弱相关	$\|0.2\|>P\geq0$	极弱相关

图 3　目标值与特征因子间的皮尔逊相关系数

Fig.3　Pearson's correlation coefficient between target values andeigenfactors

极弱相关5个等级.由图3可知,积雪覆盖率与距地表2m的气温、地表温度、露点温度、海平面气压、潜在蒸发量和净日照强度在各站点均呈现极强的相关性;与日照时数、总日照强度和直接辐射在和静、拜城站点呈现出强相关性,在其他站点呈现出极强的相关性;仅在达坂城和乌恰站点,积雪覆盖率与降雪量呈现出极弱的相关性.综合判定,各气象指标与积雪覆盖率之间存在较强的相关性,可将其视为有效的特征因子,用于构建精确的预测模型.

2.3 模型评价指标

对各模型的预测精度进行量化评估,选取均方误差(*MSE*)、决定系数(R^2)、平均绝对误差(*MAE*)和平均绝对误差百分比(*MAPE*)来评价模型的预测精度.其中,*MSE*、*MAE*、*MAPE* 越接近于0、R^2 越接近1,说明模型预测的准确性越高.具体的研究路线及流程见图4.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2 \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{13}$$

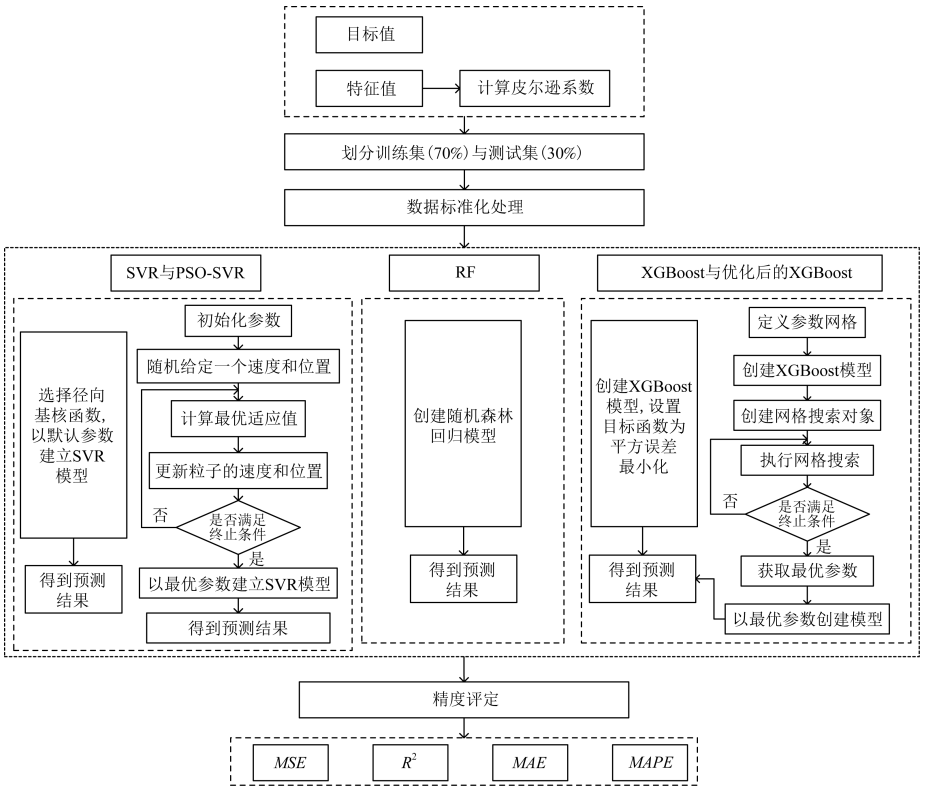


图4 技术路线

Fig.4 Technology roadmap

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{14}$$

式中:n 为样本总量;y_i 为第 i 个样本的真实值;$\hat{y}_i$ 为第 i 个样本的预测值;$\bar{y}_i$ 为样本真实值的平均值.

3 模型的建立与精度分析

3.1 SVR 及 PSO-SVR 模型的建立

基于 Anaconda 平台,在 Python3.8 环境建立各积雪覆盖率的预测模型.首先建立 SVR 模型,后利用 PSO 算法优化 SVR 模型,寻找最优惩罚因子 C 和核函数参数 g,建立 PSO-SVR 积雪覆盖率预测模型,主要流程如下:

(1)建立模型选用的特征值为气象站点的气温、气压、风速和日照时数等气象因素,因其单位不一致,首先进行特征值归一化处理;

(2)划分训练集(70%)和测试集(30%).以径向基核函数为默认参数,建立 SVR 模型,得到预测结果;

(3)利用 PSO 算法优化 SVR 模型,首先将算法参数初始化:设定最大迭代次数为 100 次,粒子种群数目为 20,惩罚因子 $C \in [0.1,100]$,核函数参数 $g \in [0.01,100]$,局部搜索能力 $c_1=2$,全局搜索能力 $c_2=2$,惯性权重因子 $\omega=0.4$,同时初始化种群粒子速度和位置;

(4)计算每个粒子的适应度.将初始化后的粒子位置向量(C,g)作为默认参数,建立 SVR 模型,预测结果的均方误差为粒子的适应度;

(5)以 20 个粒子中适应值最小的位置作为群体最优的位置;

(6)按照公式(5)、公式(6)分别更新种群粒子的速度和位置,并重复此步骤,更新优化选出种群的最小适应值,其对应粒子的(C,g)就是最优位置向量,即最优 SVR 参数;

(7)将优化后得到的最优参数赋值于 SVR 模型,建立积雪覆盖率 PSO-SVR 预测模型,得到预测结果.

将训练样本的预测结果的均方误差作为适应度函数,通过优化更新粒子的速度和位置,精准确定最优惩罚因子 C 与核函数参数 g,构建基于 PSO 算法与 SVR 的积雪覆盖率预测模型,提升预测精度与稳定性.

3.2 RF、XGBoost 和优化 XGBoost 模型的建立

3.2.1 RF 模型的建立

(1)设定参数.以 N 表示子集训练样本的数量,以 M 表示特征的总数;

(2)特征子集的选择.确定每棵决策树节点所使用的特征数目 m,其中 m 应显著小于 M;

(3)划分训练集(70%)和测试集(30%).通过有 bootstrap 抽样方式(即有放回抽样)从数据集中抽取 N 个不同的训练子集;以未被抽取的样本为验证集,用以评估模型误差;

(4)决策树的构建与分裂.对每个节点随机选取 m 个特征;基于这 m 个特征确定最佳分裂方式,以构建决策树;允许每棵树完整生长,不进行剪枝操作;

(5)集成预测.通过将多棵决策树的预测结果进行平均或加权平均,得到最终的回归结果.

3.2.2 XGBoost 模型的建立

(1)初始化.构建一棵新的回归树作为起点;

(2)计算梯度.根据最优目标函数公式,计算每个样本训练时的一阶导数和二阶导数的初始值;

(3)树生长与分割点的选择.采用近似贪心算法,寻找能够最大程度提升目标函数增益的分割点并生成新的回归树 $ft(x)$;

(4)模型集成与更新.将新生成的回归树 $ft(x)$ 加入到最终的模型中;重复执行树构建、梯度计算和模型集成的步骤,直至满足预设的收敛条件,得到 XGBoost 模型.

3.2.3 优化 XGBoost 模型的建立

(1)定义网格参数.指定在网格搜索中需要调整的参数(最大深度(*max_depth*)、树的数量(*n_estimators*)、学习率(*learning_rate*)、子样本比例(*subsample*)、每棵树使用的特征比例(*colsample_bytree*));

(2)创建 XGBoost 模型.设置目标函数为平方误差最小化,并设置随机种子以确保结果的可重复性;

(3)创建网格搜索对象.配置估算器(XGBoost 模型)、参数网格、交叉验证折数、评分机制以及详细输出级别,使用 3 折交叉验证来评估每一组参数的性能;

(4)执行网格搜索.对每一组参数组合进行模型训练和评估,以找到在交叉验证中表现最佳的参数组合.搜索完成后,以最优参数进行模型建立,得到预测结果.其中,阿勒泰站点的最优参数:*colsample_bytree* 为 0.7、*learning_rate* 为 0.05、*max_depth* 为 3、*n_estimators* 为 200、*subsample* 为 0.9.

3.3 预测结果与精度分析对比

由于选取站点较多,且各站点预测结果相差不大,仅展示阿勒泰站点(图 5)的各模型

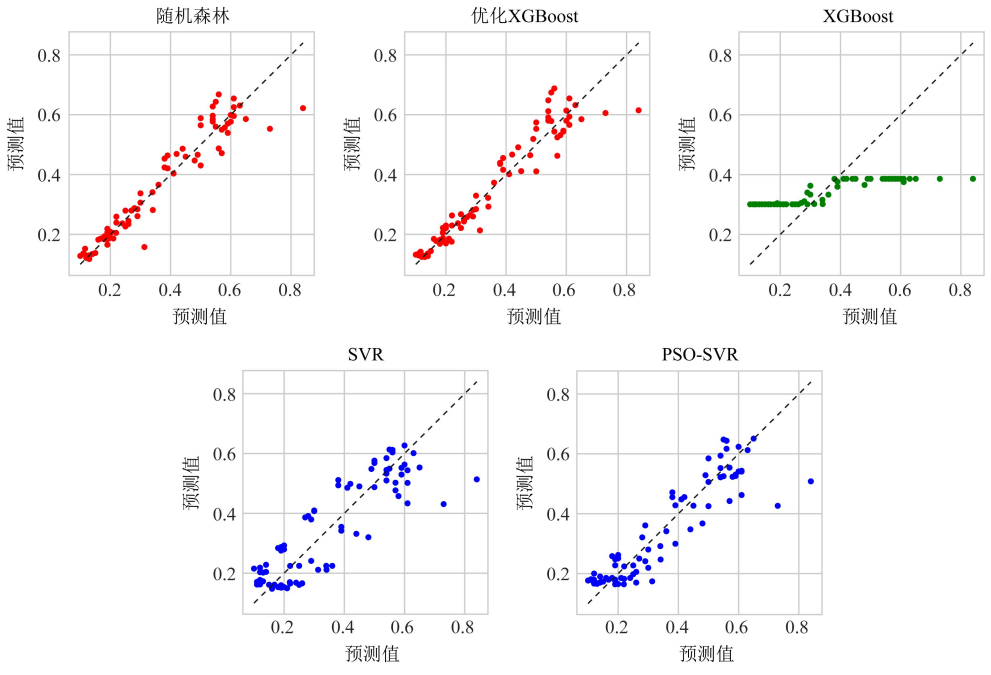


图 5 阿勒泰站点各模型预测结果

Fig.5 Predictions of each model at the Altai site

预测结果.

(1)XGBoost 模型所呈现的预测结果欠佳,预测值与真实值存在显著的误差,反映出该模型在拟合真实数据时存在局限性;

(2)RF 模型和利用网格搜索算法进行参数寻优后的 XGBoost 模型,预测数据更接近于真实数据,预测偏差较小,有着较高的预测精度;

(3)SVR 模型的预测结果虽优于 XGBoost,但其预测值与真实值仍存在一定的误差;利用 PSO 算法对 SVR 模型进行优化后,模型预测结果更为准确,但预测值与真实值的拟合程度仍低于 RF 模型与优化 XGBoost 模型.为便于更直观地对比各模型的预测准确度,绘制出各站点及 9 个站点均值的 5 个模型预测结果的 *MSE*、R^2、*MAE* 及 *MAPE* 评价指标(图 6).

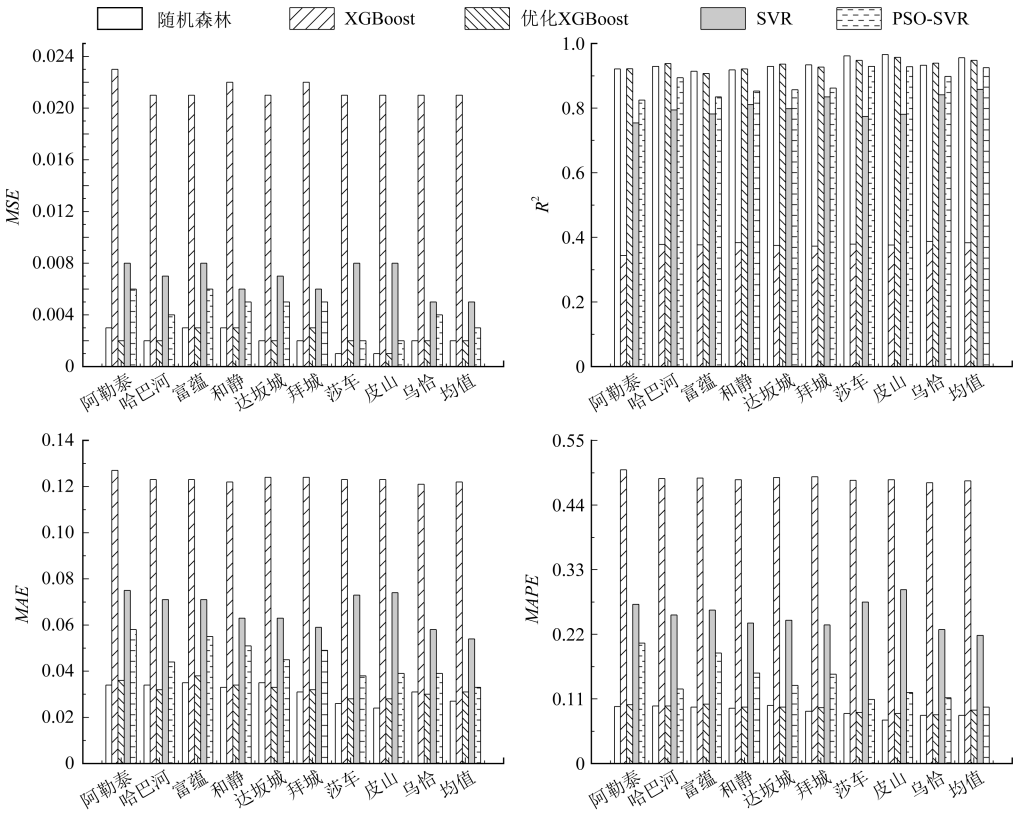


图 6　不同站点各预测模型的评价指标

Fig.6　Evaluation metrics for each predictive model at different sites

①相较于其他模型,XGBoost 模型的 *MSE*、*MAE*、*MAPE* 为最大,R^2 为最小,说明 XGBoost 在新疆积雪覆盖率的预测方面精度较低;

②RF 模型和优化 XGBoost 模型在各站点均为 $R^2>0.9$、$MSE\leqslant 0.003$、$MAE<0.04$、$MAPE<0.11$,说明 RF、优化 XGBoost 模型在新疆积雪覆盖率的预测方面呈现出较高的精度;

③SVR 模型在各站点均为 $MSE>0.004$、$MAE>0.05$、$MAPE\geqslant 0.22$,除和静、拜城、乌恰站点的 R^2 略大于 0.8 外,其余站点 R^2 均小于 0.8,说明 SVR 模型在新疆积雪覆盖率的预测方面呈现出较高精度,但低于 RF、优化 XGBoost 模型;

④PSO-SVR 模型较 SVR 模型的 *MSE*、*MAE*、*MAPE* 均有所减小，R^2 有所增大.但各站点 R^2 仍小于 RF、优化 XGBoost 模型.综合判断，5 种预测模型的精度高低等级为 RF 模型≈优化后 XGBoost 模型>PSO-SVR 模型>SVR 模型>XGBoost 模型.

4 结论

基于机器学习算法，结合新疆地势特点，分别构建各站的 SVR、PSO-SVR、RF、XGBoost 和优化 XGBoost 模型，对新疆积雪覆盖率开展预测，并对各模型进行精度分析.利用 PSO 对 SVR 进行参数寻优、利用网格搜索算法对 XGBoost 进行参数寻优，分别提高了传统 SVR 与 XGBoost 模型的预测精度，为积雪覆盖率预测领域提供了一种更加精确且便捷的新方法.得出主要结论如下：

(1)RF 与优化 XGBoost 模型在积雪覆盖率预测方面均展现出较高的预测精度；PSO-SVR 模型的预测精度略逊于前二者，XGBoost 模型的预测精度最差；SVR、XGBoost 模型优化后，对积雪覆盖率的预测准确度有所提升.这也进一步表明对传统预测模型进行参数优化是提升预测精度的重要途径；

(2)所构建的各种积雪覆盖率预测模型，其基础架构以气象因素作为主要特征因素进行考量.在实际应用中，积雪覆盖率的变化不仅仅受气象因素单一影响，而是受多种因素综合作用.其中，植被的郁闭度、种类、分布以及生长状况，通过地表遮阴、林冠截留、改变风场等方式间接影响着积雪的分配格局，太阳辐射收支平衡的影响也对积雪的消融产生影响[39]；地形地势的不同对地表积雪的积累、分布和融化速度也会产生影响，海拔通过影响温度和降水，从而影响积雪的消融与积累；坡向通过影响太阳辐射而影响积雪的消融与积累[40].此外，作为积雪的重要参数，雪深和雪线的变化不仅影响着生态系统的平衡和气候的演变，还对农业生产、交通等领域有着深远影响.在后续研究中，应考虑加入植被、地形地势等因素作为特征因子构建积雪覆盖率预测模型，以提高模型的预测能力.同时，构建雪深和雪线变化的预测模型，更好地了解积雪的变化规律，为相关领域的决策提供科学依据.

参 考 文 献

[1] 肉克亚木·艾克木，玉素甫江·如素力.伊犁河谷流域积雪分布及其变化分析[J].测绘科学，2020，45(6)：157-164
Meokyamu Aikmu，Yusufujiang Rusuli.Analysis of snowpack distribution and its changes in the Ili River Valley Basin [J].Science of Surveying and Mapping，2020，45(6)：157-164

[2] 刘 怡.新疆地区积雪、融雪型洪水与雪灾的时空变化特征研究[D].杨凌：西北农林科技大学，2020
Liu Yi.A study on the spatial and temporal variation characteristics of snowpack，snowmelt-type flooding and snowstorms in Xinjiang[D].Yangling：Northwest A&F University，2020

[3] 张庆杰，陶 辉，苏布达，等.基于 CMIP6 气候模式的新疆积雪深度时空格局研究[J].冰川冻土，2021，43(5)：1435-1445
Zhang Qingjie，Tao Hui，Su Buda，et al.Spatial and temporal pattern of snow depth in Xinjiang based on CMIP6 climate model[J].Journal of Glaciology and Geocryology.2021，43(5)：1435-1445

[4] 时兴合，李 林，陈晓光，等.青海南部牧区前冬积雪变化及其预测的关系模型研究[J].中国沙漠，2012，32(4)：1062-1070
Shi Xinghe，Li Lin，Chen Xiaoguang，et al.A relational modelling study on the change of pre-winter snowpack and its prediction in the pastoral areas of southern Qinghai[J].Journal of Desert Research，2012，32(4)：1062-1070

[5] 成 菲,李巧萍,沈新勇,等.BCC-CSM1.1m 对欧亚积雪覆盖的预测评估[J].应用气象学报,2021,32(5):553-566
Cheng Fei, Li Qiaoping, Shen Xinyong, et al. Evaluation of BCC-CSM1.1m prediction of Eurasian snow cover[J]. Journal of Applied Meteorological Science, 2021, 32(5): 553-566

[6] 郝靖宇.新疆天山山区积雪时空变化及预测分析[D].乌鲁木齐:新疆大学,2020
Hao Jingyu. Spatial and temporal variations of snow accumulation in the Tianshan Mountains of Xinjiang and analysis of forecasts[D]. Urumqi: Xinjiang University, 2020

[7] Meng Q, Ma X, Zhou Y. Forecasting of coal seam gas content by using support vector regression based on particle swarm optimization[J]. Journal of Natural Gas Science and Engineering, 2014, 21: 71-78

[8] 王龙龙,余威龙,章玉容.基于支持向量机回归的粉煤灰混凝土氯离子质量分数预测[J].浙江建筑,2024,41(3):79-83
Wang Longlong, Yu Weilong, Zhang Yurong. Prediction of chloride ion mass fraction in fly ash concrete based on support vector machine regression[J]. Zhejiang Construction, 2024, 41(3): 79-83

[9] 张永奎.支持向量机回归算法的唐山市降水量空间插值研究[J].吉林水利,2024,(2):23-25+78
Zhang Yongkui. Support vector machine regression algorithm for spatial interpolation of precipitation in Tangshan City [J]. Jilin Water Resources, 2024, (2): 23-25+78

[10] 陈家豪,郑倩茹,金立兵,等.基于 PSO-SVR 模型预测粮食孔隙率[J].粮食与油脂,2024,37(6):55-59
Chen Jiahao, Zheng Qianru, Jin Libing, et al. Prediction of grain porosity based on PSO-SVR model[J]. Cereals&Oils, 2024, 37(6): 55-59

[11] 任远芳,牛 坤,丁 静,等.基于改进 PSO 算法优化 SVR 的信息安全风险评估研究[J].贵州大学学报(自然科学版),2024,41(1):103-109
Ren Yuanfang, Niu Kun, Ding Jing, et al. Research on information security risk assessment based on improved PSO algorithm foroptimising SVR[J]. Journal of Guizhou University (Natural Science), 2024, 41(1): 103-109

[12] 刘 源,王 宇.基于随机森林的森林生态系统物候模拟研究[J].农业与技术,2024,44(11):59-62
Liu Yuan, Wang Yu. Forest ecosystem climate simulation based on random forest[J]. Agriculture and Technology, 2024, 44(11): 59-62

[13] 孙胜难,袁铸钢,刘 钊,等.基于随机森林回归算法的水泥立式磨磨内压差预测[J].洛阳理工学院学报(自然科学版),2024,34(2):44-50
Sun Shengnan, Yuan Zhugang, Liu Zhao, et al. Prediction of differential pressure in cement vertical mill based on random forest regression algorithm[J]. Journal of Luoyang Institute of Technology (Natural Science Edition), 2024, 34(2): 44-50

[14] 马赛赛,张瑞新.基于随机森林算法的露天矿抛掷爆破影响因素分析[J].露天采矿技术,2024,39(3):11-14
Ma Saisai, Zhang Ruixin. Analysis of the influence factors of cast blasting in open pit mines based on random forest algorithm[J]. Opencast Mining Technology, 2024, 39(3): 11-14

[15] 李光环,杨小天,刘训钊.XGBoost 与 GRU 模型在发电功率预测中的应用[J].福建电脑,2024,40(6):21-26
Li Guanghuan, Yang Xiaotian, Liu Xunzhao. Application of XGBoost and GRU models in power generation prediction [J]. Journal of Fujian Computer, 2024, 40(6): 21-26

[16] 曹 放,李培骏,詹同安,等.基于 XGBoost 的崩塌落石风险预测模型及在复杂山区公路工程中的应用[J].交通科技与管理,2024,5(12):1-4
Cao Fang, Li Peijun, Zhan Tongan, et al. XGBoost-based rockfall risk prediction model and its application in complex mountain highwayprojects[J]. The Technology and Management of Transportation System, 2024, 5(12): 1-4

[17] 任 伟,蒋兴浩,孙锬锋.基于 RBF 神经网络的网络安全态势预测方法[J].计算机工程与应用,2006,(31):136-138+144
Ren Wei, Jiang Xinghao, Sun Pangfeng. A network security posture prediction method based on RBF neural network[J]. Journal of Computer Research and Development, 2006, (31): 136-138+144

[18] Cortes C, Vapnik V. Suppor-vector networks[J]. Machine Learning, 1995, 20(3): 273-297

[19] 杨绪兵,陈松灿.基于原型超平面的多类最接近支持向量机[J].计算机研究与发展,2006,43(10):1700-1705
Yang Xubing, Chen Songcan. Multi-class closest support vector machine based on prototype hyperplane[J]. Journal of Computer Research and Development, 2006, 43(10): 1700-1705

[20] 刘 安.基于 PCA-SVM 模型的煤炭行业上市公司财务风险预警研究[D].宜昌:三峡大学,2021

Liu An.Research on financial risk early warning of listed companies in coal industry based on PCA-SVM model[D].Yuchang:China Three Gorges University,2021

[21] 张 驰,孙佳龙,秦江涛,等.基于支持向量回归的海洋次表层温度异常预测[J].江苏海洋大学学报(自然科学版),2020,29(2):50-57
Zhang Chi,Sun Jialong,Qin Jiangtao,et al.Prediction of oceanic subsurface temperature anomalies based on support vectorregression[J].Journal of Jiangsu Ocean University (Natural Science Edition),2020,29(2):50-57

[22] 周裕群,张德生,张 晓.一种改进的鲁棒模糊孪生支持向量机算法[J].计算机工程与应用,2023,59(1):140-148
Zhou Yuqun,Zhang Desheng,Zhang Xiao.An improved robust fuzzy twin support vector machine algorithm[J].Journal of Computer Research and Development,2023,59(1):140-148

[23] Smola A J,Schölkopf B.A tutorial on support vector regression[J].Statistics and Computing,2004,14:199-222

[24] 孙玉婷,王映龙,杨红云,等.基于支持向量机回归预测水稻叶片 SPAD 值[J].科技通报,2018,34(9):55-59
Sun Yuting,Wang Yinglong,Yang Hongyun,et al.Predicting SPAD values of rice leaves based on support vector machine regression[J].Bulletin of Science and Technology,2018,34(9):55-59

[25] Cherkassky V,Ma Y.Practical selection of SVM parameters and noise estimation for SVM regression[J].Neural Networks,2004,17(1):113-126

[26] 杨 栩,尤学一,季 民.天津城市绿地土壤水分特征曲线模型及参数确定[J].干旱区资源与环境,2013,27(8):115-119
Yang Xu,You Xueyi,Ji Min.Modelling and parameter determination of soil moisture profile in urban green areas of Tianjin[J].Journal of Arid Land Resources and Environment,2013,27(8):115-119

[27] Maihemuti S,Wang W,Wang H,et al.Voltage security operation region calculation based on improved particle swarm optimization and recursive least square hybrid algorithm[J].Journal of Modern Power Systems and Clean Energy,2020,9(1):138-147

[28] 秦文静,樊贵盛.基于粒子群优化算法-支持向量机的原状黄土 Van Genuchten 模型参数土壤传输函数[J].干旱区资源与环境,2020,34(11):132-137
Qin Wenjing,Fan Guisheng.Parameter soil transfer function of Van Genuchten model for primary loess based on particle swarm optimisation algorithm-support vector machine[J].Journal of Arid Land Resources and Environment,2020,34(11):132-137

[29] 高佳南,吴奉亮,马 砺,等.矿井淋水井筒风温 PSO-SVR 预测方法[J].西安科技大学学报,2022,42(3):476-483
Gao Jianan,Wu Fengliang,Ma Li,et al.PSO-SVR prediction method of wind temperature in mine drench shaft[J].Journal of Xi'an University of Science and Technology,2022,42(3):476-483

[30] 张范平,唐德善,戴会超,等.基于 CPSO 参数辨识的支持向量机增泄水量计算模型研究[J].干旱区资源与环境,2014,28(12):117-121
Zhang Fanping,Tang Deshan,Dai Huichao,et al.Research on support vector machine computational model for water augmentation and discharge based on CPSO parameter identification[J].Journal of Arid Land Resources and Environment,2014,28(12):117-121

[31] Breiman L.Random forests[J].Machine Learning,2001,45:5-32

[32] 赵华生,金 龙,黄小燕,等.基于 CNN 和 RF 算法的 ECMWF 降水分级订正预报方法[J].气象科技,2021,49(3):419-426
Zhao Huasheng,Jin Long,Huang Xiaoyan,et al.A hierarchical revised forecast method for ECMWF precipitation based on CNN and RF algorithms[J].Meteorological Science and Technology,2021,49(3):419-426

[33] 许垚涛,李 鹏,马方铭,等.不同机器学习模型在流域输沙模拟中的应用与解释[J/OL].应用基础与工程科学学报,1-14[2024-10-12]
Xu Yaotao,Li Peng,Ma Fangming,et al.Application and interpretation of different machine learning models in watershed sand transport simulation[J/OL].Journal of Basic Science and Engineering,1-14[2024-10-12]

[34] Chen T,Guestrin C.Xgboost:A scalable tree boosting system[C].Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:785-794

[35] 王迎超,郭 崟,姜 雯,等.基于 XGBoost 算法的公路隧道失稳风险评估模型及系统开发[J].应用基础与工程科学学报,2024,32(4):957-971
Wang Yingchao,Guo Yin,Jiang Wen,et al.Risk assessment model and system development of road tunnel instability based on XGBoost algorithm[J].Journal of Basic Science and Engineering,2024,32(4):957-971

[36] 薛 强,吕继强,罗平平,等.和田河流域山区积雪覆盖时空变化规律研究[J].中国农村水利水电,2020,(1):88-96
Xue Qiang,Lü Jiqiang,Luo Pingping,et al.Study on spatial and temporal variation rules of snow cover in mountainous areas of Hotan River Basin[J].China Rural Water and Hydropower,2020,(1):88-96
[37] 叶聪霄.青藏高原积雪深度时空变化及其影响因素分析[D].南京:南京信息工程大学,2023
Ye Congxiao.Analysis of spatial and temporal variations of snow depth and its influencing factors on the Tibetan Plateau [D].Nanjing:Nanjing University of Information Science & Technology,2023
[38] Leathers D J,Robinson D A.Abrupt changes in the seasonal cycle of north American snow cover[J].Journal of Climate,1997,10(10):2569-2585
[39] 王计平,蔚奴平,丁 易,等.森林植被对积雪分配及其消融影响研究综述[J].自然资源学报,2013,28(10):1808-1816
Wang Jiping,Wei Nuping,Ding Yi,et al.A review on the effects of forest vegetation on snow distribution and its ablation [J].Journal of Natural Resources,2013,28(10):1808-1816
[40] 李 虹,李忠勤,陈普晨,等.近 20a 新疆阿尔泰山积雪时空变化及其影响因素[J].干旱区研究,2023,40(7):1040-1051
Li Hong,Li Zhongqin,Chen Puchen,et al.Spatial and temporal variations of snowpack in the Altai Mountains of Xinjiang in the last 20a and their influencing factors[J].Arid Zone Research,2023,40(7):1040-1051

Construction of a Snow Cover Prediction Model in Xinjiang Based on Machine Learning Algorithm

DENG Wenbin, HOU Xueqing
(College of Architecture and Engineering,Xinjiang University,Urumqi 830046,China)

Abstract

Snow cover is a kind of valuable freshwater resources,and the change of snow coverage rate has a profound impact on the development of agriculture and animal husbandry.There has been little research on the prediction of such coverage rates so far.In order to improve the accuracy of such prediction,this study uses machine learning algorithms to construct Support Vector Regression (SVR),PSO-optimized SVR,Random Forest (RF),XGBoost and optimized XGBoost prediction models,which are utilized to predict snow cover in Xinjiang,while comparing and analyzing the prediction accuracy of the models.The results show that the R^2 values of both RF and optimized XGBoost models are greater than 0.9;the R^2 values of all traditional SVR models is less than 0.8;the R^2 values of all PSO-optimized SVR models are greater than 0.8,with some greater than 0.9;and the R^2 values of all XGBoost models are lower than 0.4.These data indicate that the RF,optimized XGBoost and PSO-SVR models can deliver high prediction accuracy for snow cove;the XGBoost models have the poorest prediction results;and it is necessary to optimize traditional models with different algorithms.

Keywords:snow cover;support vector machine regression;particle swarm optimization algorithm;RF;XGBoost;parameter optimization