引用格式：毛亚萍,房世峰.基于机器学习的参考作物蒸散量估算研究[J].地球信息科学学报,2020,22(8):1692-1701. [Mao Y P, Fang S F. Research of reference evapotranspiration's simulation based on machine learning[J]. Journal of Geo-information Science, 2020,22(8):1692-1701.] **DOI:**10.12082/dqxxkx.2020.200085

基于机器学习的参考作物蒸散量估算研究

毛亚萍[1,2],房世峰[1*]

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101；2. 中国科学院大学,北京 100049

Research of Reference Evapotranspiration's Simulation based on Machine Learning

MAO Yaping[1,2], FANG Shifeng[1*]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Accurate estimation of Reference Evapotranspiration (ET_0) is essential to agricultural water management and allocation and hydrological cycle research. FAO-56 Penman-Monteith (FAO-56 PM) is the standard method to calculate ET_0 recommended by Food and Agriculture Organization of the United Nations (FAO). But this method demands too many parameters and these meteorological inputs are not commonly available or unreliable, especially in Xinjiang province. Under this situation, machine learning algorithms have been introduced to estimate ET_0 using fewer meteorological parameters and many comparisons of their prediction accuracy have been conducted. But the input combinations of meteorological factors are various and lack theoretical support. Meanwhile, the comparison of their performance at different time-scales has not been comprehensively conducted yet, and the good stability and less computational effort of models are also less to consider. The objective of this research was to evaluate machine learning algorithms' performance in modeling daily ET_0 and monthly ET_0 using fewer meteorological factors in Xinjiang. At this point, by using data collected from 41 weather stations in Xinjiang, this paper used Sensitivity Coefficient (S_V) to evaluate the meteorological factors' influence degree to ET_0 and then combined factors with high influence as input to Support Vector Machine (SVM), Gradient Boosted Decision Tree (GBDT), Random Forest (RF), and Extreme Learning Machine (ELM) in modeling daily and monthly ET_0, and finally investigated and compared the performance of these algorithms from accuracy, stability and computational cost. The results showed RH (S_I=−0.516), T_{max} (S_I=0.283) and U_2 (S_I=0.266) had high influence to ET_0 followed by n (S_I=0.124), while T_{min} (S_I=−0.016) and T_{avg} (S_I=−0.003) exhibited low influence. In modeling daily ET_0, models obtained satisfactory accuracy ($RMSE$<0.5 mm/day, R^2>0.95) with input combination of RH, T_{max}, U_2 and n, while combination of RH、T_{max} and U_2 showed comparable accuracy for monthly ET_0 prediction. The SVM and GBDT models showed the best performance and

收稿日期：2020-02-18；**修回日期**：2020-04-16.

基金项目：中国科学院战略性先导科技专项(XDA20010302)；国家自然科学基金项目(41971082、U1503184)。[**Foundation items:** Strategic Priority Research Program of the Chinese Academy of Sciences, No.XDA20010302; National Natural Science Foundation of China, No.41971082, U1503184.]

作者简介：毛亚萍(1992—)，女，四川新津人，硕士生，主要从事环境物联网及其应用研究。E-mail: maoyp.17s@igsnrr.ac.cn

***通讯作者**：房世峰(1982—)，男，湖北宜昌人，博士，高级工程师，硕士生导师，主要从事生态环境物联网及陆面过程模型模拟研究。E-mail: fangsf@igsnrr.ac.cn

have been recommended for daily and monthly ET_0 estimation in Xinjiang and maybe elsewhere with similar climates around the world.

Key words: Reference Evapotranspiration; machine learning; Penman-Monteith; Xinjiang; Support Vector Machine; Random Forest; Gradient Boosted Decision Tree; Extreme Learning Machine

***Corresponding author:** FANG Shifeng, E-mail: fangsf@igsnrr.ac.cn

摘要：参考作物蒸散量（Reference Evapotranspiration, ET_0）的准确估算对区域水资源管理和分配、流域水量平衡以及气候变化等研究具有重要作用。新疆地处我国西北干旱地区，水资源供需矛盾尖锐，精确估算该地区的ET_0有助于其科学合理地调配水资源，缓解水资源供需压力。FAO推荐的Penman-Monteith法是计算ET_0的标准方法，但该方法需要多项气象因子，而新疆地区气象站点较少且分布不均，精确完备的气象数据在新疆大部分区域难以获取。因此，如何使用有限气象因子获取高精度的ET_0在新疆地区备受关注。本文基于中国气象数据网提供的新疆地区1980—2019年的地面气候资料日值数据集，在日和月尺度下，通过对最高气温T_{max}、最低气温T_{min}、平均气温T_{avg}、风速U_2、相对湿度RH和日照时数n共6项气象因子进行敏感性分析，形成不同的气象因子组合；然后使用SVM、RF、GBDT和ELM 4种机器学习算法，以FAO-56 PM计算值为标准值，对新疆地区的ET_0进行了拟合建模；最后，从拟合精度、稳定性和计算代价3个方面对模型进行评价。研究表明：① 在新疆地区，ET_0对RH、T_{max}和U_2敏感系数级别为高，平均敏感系数分别为−0.516、0.283和0.266；n为中等，平均敏感系数为0.124；T_{min}和T_{avg}为低，平均敏感系数分别为−0.016和−0.003；② 在日尺度，各算法在RH、T_{max}、U_2和n这4项气象因子为输入时精度较高（$RMSE$<0.5 mm/day，R^2>0.95），可对ET_0进行精确估算；在月尺度，各算法使用RH、T_{max}和U_2这3项参数便可对ET_0进行精确估算。SVM和GBDT这2种算法在日尺度和月尺度都有较好的适用性，可在相应尺度下使用较少气象因子替代FAO-56 PM公式对ET_0进行估算。

关键词：参考作物蒸散量；机器学习；Penman-Monteith；新疆；支持向量机；随机森林；梯度提升树；极限学习机

1 引言

参考作物蒸散量（Reference Evapotranspiration, ET_0）是计算作物蒸散量的关键因子[1]，其准确估算对水资源管理和分配、流域水量平衡以及气候变化等研究具有重要作用。新疆地处我国西北干旱地区，水资源供需矛盾尖锐，精确估算该地区的ET_0有助于其科学合理地调配水资源，缓解水资源供需压力。目前，ET_0标准计算方法是由联合国粮食与农业组织（Food and Agriculture Organization of the United Nations, FAO）推荐的Penman-Monteith（FAO-56 PM）法。该方法将能量平衡和空气动力理论相结合，在不同区域和不同气候条件下都有着极强的适用性[1]，但需要气温、相对湿度、太阳辐射和风速等多项气象因子。然而，新疆地区气象站点较少且分布不均，导致精确完备的气象因子在新疆大部分区域难以获取。因此，如何使用有限气象因子获取高精度的ET_0在新疆地区备受关注。

近年来，随着数据挖掘技术的兴起，不少学者基于机器学习使用有限的气象因子对ET_0进行了估算研究，并获得了精度较高的模型。目前，拟合精度较高的机器学习算法主要有3类：① 核函数算法，如支持向量机（Support Vector Machine, SVM）[2]，KNEA（Kernel-based Arps Decline Model）[3]等；② 树的集成算法，如随机森林算法（Random Forest, RF）[4]，梯度树提升（Gradient Boosted Decision Tree, GBDT）[5-6]和极限梯度提升树（Extreme Gradient Boosting, XGBoost）[7]等；③ 神经网络算法，如广义回归神经网络（Generalized Regression Neural Networks, GRNN）[8-10]，极限学习机（Extreme Learning Machine, ELM）[9,11-14]等。王升等[4]和张皓杰等[11]在最高气温、最低气温、风速、相对湿度和日照时数等气象因子的15种不同组合下，分别将RF和ELM 2种机器学习算法，与传统基于温度的Hargreaves方法进行了对比分析，发现RF和ELM模型在仅以气温作为输入时，精度仍高于Hargreaves。Wu等[3]在中国7个不同气候地区，将SVM与其他7种机器学习算法进行了对比，发现SVM在各地区都具有较高的拟合精度。Reis等[15]考虑数据分割模式对算法拟合精度的影响，在单个气象站和混合多站数据的2种数据模式下，评估了ANN（Artificial Neural Network）、PR（Polynomial Rgression）、ELM及Hargreaves 4种模型对日ET_0的拟合效果，发现ANN、PR和ELM 3种机器学习模型拟合精度都高于Hargreaves模型，且在单个站点的拟合精度高于混合多站的拟合精度。然而，截至目前，在基于机器学习使用有限气象因子对ET_0的估算研究中，对不同时间尺度下算法拟合效果的对比研究较少，且存在气象因子组合较为盲

目，导致气象因子组合过多，以及模型评价指标较为片面（只以精度作为评价指标）等缺点。

因此，为通过有限气象因子获取新疆地区高精度的ET_0，本文基于中国气象数据网提供的新疆地区40年（1980—2019）的地面气候资料日值数据集，在日和月尺度下，首先，通过对气象因子进行敏感性分析，形成不同的气象因子组合；然后分别选取了目前拟合ET_0精度较高的核函数算法SVM，树的集成算法RF和GBDT，以及神经网络算法ELM共4种算法，以FAO-56 PM计算值为标准值，对新疆地区的ET_0进行了拟合建模；最后，从拟合精度、稳定性和计算代价3个方面对模型进行评价，以期获得使用较少气象因子而综合表现较优的ET_0拟合模型，为该地区水资源管理提供科学依据。

2 研究区概况及数据来源

2.1 研究区概况

本文以新疆维吾尔自治区为研究区，地理坐标范围为34°20′ N—40°13′ N，73°28′ E—96°24′ E。整体气候为温带大陆性气候，内部以天山为分界线，可分为南北两部分。北疆为温带大陆性干旱半干旱气候区，年均气温−4~9 ℃，全年降水大于150 mm；南疆为暖温带大陆性干旱气候区，年均气温7~14 ℃，全年降水量小于100 mm。

2.2 数据情况

2.2.1 数据预处理

本文数据来源于中国气象数据网提供的地面气候资料日值数据集，数据时间跨度从1980年1月1日至2019年12月31日，包含了新疆地区41个气象站每日的平均气温T_{avg}、最高温度T_{max}、最低温度T_{min}、平均风速U、平均相对湿度RH和日照时数n等计算ET_0所需参数，气象站点分布情况如图1所示。由于数据集中缺失数据占比仅为0.3%，因此舍弃了缺失数据。

月值数据由日值数据平均整合得到，其中U在整合前，已根据FAO于1998年发布的作物需水量计算指南[1]（FAO-56指南），转换为地面以上2 m标准高度下测量的风速U_2，转换公式为：

$$U_2 = \frac{4.87U}{\ln(67.8z - 5.42)} \quad (1)$$

式中：U_2为地面以上2 m标准高度下测量的风速；U为原平均风速；z为原风速测量高度，根据地面气象观测规范，在本文中其值为10 m。

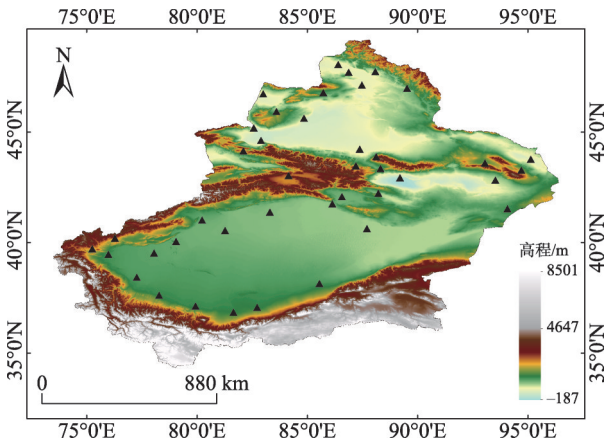


图1 新疆区域41个气象站点分布情况

Fig.1 The geographical locations of the forty-one stations in Xinjiang

2.2.2 各尺度数据

在月尺度，本文选取新疆全区域41站点作为数据来源站点，数据量约为19 680条；而在日尺度，为避免数据量和数据区域分布的差异性对模型拟合结果产生影响，本文分别从每个站点随机抽取约480条数据，舍弃其他数据，形成了包含19 742条数据的数据集。各尺度数据情况如表1所示。

表1 1980—2019年新疆日尺度和月尺度下地面气候数据集参数均值统计

Tab. 1 Climatic data averages at daily scale and monthly scale in Xinjiang from 1980 to 2019

尺度	数据量/条	平均值(1980-01-01—2019-12-31)					
		T_{avg}/℃	T_{max}/℃	T_{min}/℃	U_2/m·s^{-1}	RH/%	n/h
日	19 742	8.09	14.98	2.08	1.68	51.49	7.76
月	19 680	8.53	15.42	2.49	1.69	51.16	7.83

3 研究方法

3.1 技术路线

本文技术路线如图2所示，首先需做3个方面的准备：① 通过对气象因子进行敏感性分析，形成不同的气象因子组合，作为输入参数；② 使用FAO-56 PM公式计算ET_0，作为算法训练的标准值；③ 对数据集进行处理，分割为训练集和测试集。然后，使用训练集对SVM、RF、GBDT和ELM 4种机器学习算法进行训练，获得相应的模型。最后，本文将从精度、稳定性和计算代价3个方面对模型进行评价。

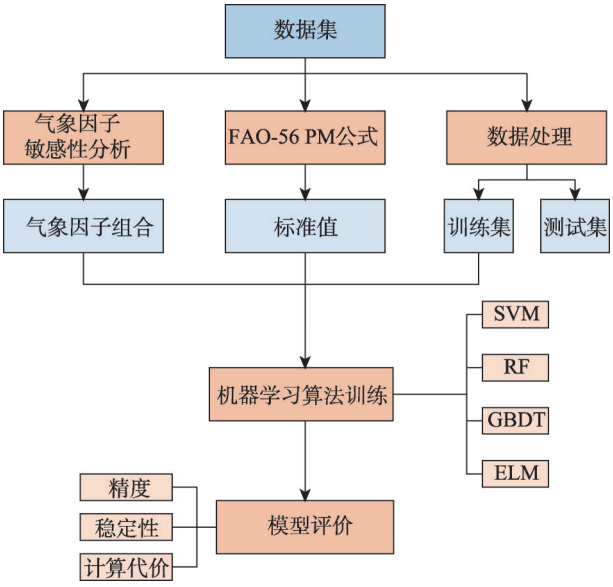


图2　本研究的技术路线

Fig.2　Methodological workflow of this study

3.2　气象因子敏感性分析

为避免气象因子组合过多，本文使用敏感系数对气象因子进行敏感性分析，剔除对ET_0影响较小的气象因子，形成不同的气象因子组合，作为机器学习算法训练时的输入参数。敏感系数是表征气象因子对ET_0趋势变化产生影响的定量参数，定义为ET_0与气象因子变化率之比[16]：

$$S_{vi}=\lim_{\Delta v_i\to 0}\left(\frac{\Delta ET_0/ET_0}{\Delta v_i/v_i}\right)=\frac{\partial ET_0}{\partial v_i}\cdot\frac{v_i}{ET_0} \tag{2}$$

式中：S_{vi}为气象因子v_i的敏感系数；v_i、Δv_i为气象因子和气象因子变化量；ΔET_0为ET_0随气象因子变化而产生的变化量。S_{vi}为正值，表示气象因子与ET_0变化趋势相同，反之，则变化趋势相反；S_{vi}绝对值越大，表示对ET_0影响越大。为更好地评价气象因子的敏感性，本文引入了Lenhart等[17]对敏感系数的分级方法，如表2所示。

3.3　FAO-56Penman-Monteith（FAO-56PM）公式

本文以FAO-56 PM法计算的ET_0为标准值，对机器学习算法进行训练和评价，具体计算公式为：

$$ET_0=\frac{0.408\Delta(R_n-G)+\gamma\dfrac{900}{T+273}U_2(e_s-e_a)}{\Delta+\gamma(1+0.34U_2)} \tag{3}$$

式中：ET_0为参考作物蒸散量/(mm/day)；R_n是作物表面上的净辐射/(MJ/(m^2· day))；G是土壤热通量/(MJ/(m^2· day))，在日尺度下可忽略G对ET_0的影响，在月尺度需根据月平均气温进行估算；T是2 m高处日平均气温/℃；U_2为2 m高处风速/(m/s)；e_s为饱和水汽压/kPa；e_a为实际水汽压/kPa；Δ是饱和水汽压曲线的倾率；γ是湿度计常数。根据FAO指南，以上参数可由表3中参数直接或间接得到。

表3　FAO-56 PM公式所需参数

Tab. 3　The parameters demanded by the FAO-56 PM equation

参数类别	具体参数
日期	年内日序数J
地理位置参数	高程z/m、纬度φ
气象参数	日最高气温T_{max}/℃、日最低气温T_{min}/℃、日平均气温T_{avg}/℃、日平均风速U_2/(m/s)、日平均相对湿度RH/%、日照时数n/h

3.4　数据处理

3.4.1　数据标准化

为消除不同气象因子之间量纲对算法拟合效果的影响，本文对数据集进行了归一化，将各因子的值映射到[0, 1]区间，公式为：

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{4}$$

式中：x_{max}和x_{min}分别为数据集中各因子的最大值和最小值；x和x'分别为输入值和归一化后的值。

3.4.2　数据分割与交叉验证

本文以3∶1的比例将数据随机分割为训练集（75%）和测试集（25%），然后在训练集进行了5折交叉验证，得到模型参数；最后用测试集对模型精度进行评估。

3.5　机器学习算法

目前，拟合精度较高的机器学习算法主要有核函数的算法、树的集成算法和神经网络算法3类，本文从这3类算法中，分别选取了拟合ET_0精度较高核函数算法SVM，树的集成算法RF和GBDT，以及神经网络算法ELM4种算法对ET_0进行拟合建模。

3.5.1　支持向量SVM（高斯核）

SVM是一种二分类模型，其基本模型是定义在

表2　敏感系数分级

Tab. 2　Classification of the sensitivity coefficient

敏感系数范围	$0.00\leqslant\lvert S_{vi}\rvert<0.05$	$0.05\leqslant\lvert S_{vi}\rvert<0.20$	$0.20\leqslant\lvert S_{vi}\rvert<1.00$	$\lvert S_{vi}\rvert\geqslant 1.00$
级别	低	中等	高	非常高

特征空间上的间隔最大化的线性分类器，但通过引入核函数，其可变为非线性分类器。当SVM用于回归时，其算法目标是找到一个平面，使数据集中所有点到平面的距离和为最小，其算法目标函数可以表示为[18]：

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}l_\varepsilon\left(f\left(x_i\right)-y_i\right) \quad (5)$$

式中：n为样本数量；$f(x_i)$为拟合结果，$f(x_i)=w^Tx+b$，w为平面法向量，b为模型参数；y_i为目标值；C惩罚系数，C越大，对损失的惩罚越大，C越小，对损失惩罚越小。本研究中对C以1步长在[1, 10]的区间进行网格搜索。$l_\varepsilon(e)$为带软间隔的损失函数，表达式为：

$$l_\varepsilon(e)=\begin{cases}0 & if|e|\leqslant\varepsilon\\ |e|-\varepsilon & 其他\end{cases} \quad (6)$$

式中：e为误差；ε为模型对误差的容忍值，即当拟合值$f(x_i)$与y_i的差值小于ε时，不计为损失，本文中ε值设为0.1。

根据Kisi等[19]的研究结果，本研究采用了在拟合ET_0时表现较优的高斯核函数对ET_0进行非线性拟合，其表达式为：

$$k(x)=\exp\left(-\frac{\|x-l\|^2}{2\sigma^2}\right) \quad (7)$$

式中：x为训练数据，l为人工选择的中心点，δ为高斯核带宽，本研究中分别设置为(0，0)和1。

3.5.2 随机森林RF

RF是由Breiman[20]在2001年提出的一种基于决策树的集成算法，决策树有关理论可参考李航[21]相关研究。当RF用作回归时，其基本理论是通过有放回的随机抽取样本，生成多个决策树，然后将多棵树的决策结果进行平均作为最终输出。其特点是在生成树时，每个节点抽取的样本，在特征个数和数据量上都会随机化。因此，RF具有能够评估特征重要性、处理高维特征数据等优点[22]。

RF具有2个重要的自定义参数：树的数量和分割节点的特征数量。在本文中由于已对气象因子进行敏感性分析，因此分割节点的特征数量设为各组合下气象因子的个数；而对树的数量，本文以10为步长，在[10, 100]的区间进行搜索。

3.5.3 梯度提升树GBDT

GBDT是由Friedman[23]在2002年提出的算法，其原理与RF类似，都是基于树的集成算法。区别在于GBDT树是依次构建的，即第一棵树对所有样本进行训练，得到一个模型及其权值；后一颗树以减少前一颗树的残差为目标，对所有样本进行训练，得到模型和权值；当残差足够小或者达到设置的树的数量时则停止，最终的模型是将每颗树的结果加权求和得到，因此GBDT能够减少偏差。与RF相比，GBDT计算代价更小且更不容易过拟合。GBDT的具体计算过程可参考Friedman相关研究，在本文中GBDT的参数搜索方法与RF相同。

3.5.4 极限学习机ELM

ELM是由黄广斌[24]在2004年提出的一种对单隐藏层前馈神经网络(Single-hidden Layer Feedforward Neural Network, SLFN)的优化算法。传统SLFN采用梯度下降法求解输入层的权值矩阵w和偏置b，而ELM则先随机选取w和b，然后以最小化输出值与目标值的矩阵为目标，通过广义逆矩阵理论解析求解w和b。与传统SLFN相比，ELM具有训练参数少、学习速度快和泛化能力强的优点，其目标函数可以表示为：

$$\min\|\boldsymbol{H\beta}-\boldsymbol{T}\|^2,\quad \boldsymbol{\beta}\in\mathrm{R}^{L\times p} \quad (8)$$

式中：$\boldsymbol{H}$为隐藏层输出矩阵；$\boldsymbol{T}$为目标值矩阵；$\boldsymbol{\beta}$为输出层权重矩阵；R表示有理数集；L为隐藏层节点个数；p为输出层节点个数，$p\geqslant1$。$\boldsymbol{H}$与$\boldsymbol{T}$可表示为：

$$\boldsymbol{H}=\begin{bmatrix}h_1(x_1) & \cdots & h_L(x_1)\\ \cdots & \cdots & \cdots\\ h_1(x_n) & \cdots & h_L(x_n)\end{bmatrix} \quad (9)$$

$$\boldsymbol{T}=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix} \quad (10)$$

式中：n为训练数据量，$h_i(x)$为隐藏层激活函数。在本文中，隐藏层神经元个数以1为步长，在[5, 10]范围内依次搜索；激活函数则分别搜索了Sigmoid、Tanh和ReLU共3种数学函数。

3.6 模型评价方法

本文从精度、稳定性和计算代价3个方面对机器学习算法训练得到的模型进行评价，精度表征了模型拟合的精确度；稳定性能够表征模型是否存在过拟合或欠拟合现象；计算代价表征了模型的复杂程度。

3.6.1 精度

(1)决定系数R^2

决定系数(Coefficient of Determination，COD)是评价回归模型系数拟合优度的常用指标，R^2越大，模型拟合结果越准确，其计算公式为：

$$R^2=\frac{\left[\sum_{i=1}^{N}(y_i-\bar{y})(y_i^{'}-\overline{y^{'}})\right]^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2\sum_{i=1}^{N}(y_i^{'}-\overline{y^{'}})^2} \tag{11}$$

式中：y_i是FAO-56 PM公式计算值；$y_i^{'}$是机器学习算法拟合值，$\bar{y}$、$\overline{y^{'}}$分别是y_i、$y_i^{'}$的平均值；N为数据集的数据量。

（2）均方根误差RMSE

均方根误差（Root Mean Squared Error，RMSE）是评价回归模型拟合结果与目标值差别大小的指标，均方根误差越小，模型拟合效果越好，其计算公式为：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y_i-y_i^{'})^2}{N}} \tag{12}$$

3.6.2 稳定性

对比模型在训练集和测试集的精度变化，可评价模型的稳定性。若模型在训练集精度很高，而在测试集精度较差，则模型存在过拟合现象，稳定性差；若模型在训练集和测试集的精度变化不大，则模型较为稳定；若模型在训练集和测试集精度都较差，则模型存在欠拟合现象。

3.6.3 计算代价

计算代价是指算法通过训练得到模型所用的时间，计算代价越高表示模型越复杂，耗时高；反之，模型较为简单，耗时低。

4 结果与分析

4.1 气象因子敏感性分析

日尺度和月尺度数据集中各气象因子敏感系数如表4所示，可见在2种尺度下，气象因子敏感性表现一致。6项因子对ET_0影响由高到低分别为RH、T_{max}、U_2、n、T_{min}和T_{avg}，其中RH、T_{max}和U_2敏感系数级别为高，n为中等，T_{min}和T_{avg}为低。在新疆地区，RH、T_{min}和T_{avg}的敏感系数为负值，其值的增加将导致ET_0的减小；T_{max}、U_2和n的敏感系数为正值，其值增加ET_0也随之增加。

表4　日尺度和月尺度下气象因子敏感系数

Tab. 4　Sensitivity coefficients of meteorological parameters at daily and monthly scale

尺度	T_{avg}	T_{max}	T_{min}	U_2	RH	n
日	−0.001	0.28	−0.011	0.266	−0.529	0.107
月	−0.005	0.286	−0.021	0.267	−0.503	0.141

根据表4结果，本文选取了T_{max}、U_2、RH和n 4项气象因子形成了4不同的气象因子组合，使用机器学习算法对ET_0进行拟合建模。气象组合具体情况如表5所示。

表5　气象因子的不同组合

Tab. 5　Input combinations of meteorological parameters

组合	输入参数
Group 1	RH、T_{max}
Group 2	RH、U_2
Group 3	RH、T_{max}、U_2
Group 4	RH、T_{max}、U_2、n

4.2 精度分析

分别以表5中的气象因子组合为输入参数，通过训练得到模型后，各模型在测试集的精度如表6所示。可见，各算法所得模型的精度随气象因子输入个数的增多而提高，在各组合取得的精度由高到低分别为Group4、Group3、Group1和Group2。与有两项参数的Group1和Group2相比，Group1的精度远大于Group2的精度，以SVM为例，在日尺度和月尺度上，其Group1的*RMSE*比Group2分别降低了43.6%和51.1%，R^2提高了68%和56.3%。

在日尺度上，各算法在Group4取得的精度较优（*RMSE*<0.5mm/day，R^2>0.95）；在月尺度上，各算法在仅有3个气象因子的Group3便取得了较优的精度。同时，各算法在月尺度上的精度整体高于日尺度，以SVM为例，其月尺度的*RMSE*在Group1—Group4比日尺度分别降低了25.8%、14.3%、49.3%和39.0%，R^2分别增加了7.0%、15.0%、3.7%和1.6%。这主要因为月尺度数据是由日尺度数据平均得到，与日尺度数据相比，数据的特征分布被缩小，更容易得到精度高的模型。

在日尺度上，SVM模型在Group1、Group3和Group4时SVM精度最高；在Group2时，与最高精度相比，其*RMSE*仅增加了2%，R^2降低了3.5%，总体表现最优。GBDT在Group2时，精度最高；在Group1和Group3与SVM精度相当（*RMSE*差值<0.02 mm/day）；在Group4与SVM相比，*RMSE*增加了8.4%，R^2降低了0.5%，总体表现次优。ELM在Group1和Group2上分别与SVM和GBDT取得了相当的精度；在Group3和Group4与SVM相比，*RMSE*分别增加了4.4%和10.7%，R^2分别降低了0.4%和0.6%，总体表现较差。RF在Group4与SVM

表6 日尺度和月尺度下机器学习模型在测试集的精度

Tab. 6 The accuracy of machine learning models ontest set at daily and monthly scale

算法	气象组合	*RMSE*/(mm· day^{-1})		R^2	
		日尺度	月尺度	日尺度	月尺度
SVM	Group1	0.934	0.693	0.840	0.899
	Group2	1.655	1.418	0.500	0.575
	Group3	0.521	0.264	0.950	0.985
	Group4	0.392	0.239	0.972	0.988
RF	Group1	1.076	0.736	0.788	0.885
	Group2	1.773	1.500	0.426	0.524
	Group3	0.563	0.270	0.942	0.985
	Group4	0.412	0.241	0.969	0.988
GBDT	Group1	0.939	0.680	0.839	0.902
	Group2	1.624	1.380	0.518	0.597
	Group3	0.541	0.275	0.947	0.984
	Group4	0.425	0.257	0.967	0.986
ELM	Group1	0.947	0.710	0.836	0.893
	Group2	1.625	1.411	0.517	0.579
	Group3	0.544	0.290	0.946	0.982
	Group4	0.434	0.274	0.966	0.984

精度相当;但在Group1—Group3,与最高精度相比,*RMSE*分别增加了9.2%、8.1%和5.1%,R^2分别降低了6.2%、17.7%和0.8%,总体表现最差。在月尺度上,SVM与GBDT算法在所有组合的表现相当(*RMSE*差值<0.02 mm/day),且为最优。RF在Group3和Group4与最高精度相当,但在Group1和Group2表现相对较差,与最高精度相比,其*RMSE*分别增加了8.2%和8.7%,R^2分别降低了1.9%和12.2%。ELM在Group1和Group2精度高于RF,但在Group3和Group4精度远低于RF,*RMSE*分别增加了7.4%和13.7%,R^2分别降低了0.3%和0.4%,总体表现最差。

4.3 稳定性分析

各尺度下,各模型在训练集和测试集取得的*RMSE*柱状图如图3所示,可见SVM和ELM在各尺度及各气象因子组合下,*RMSE*变化率基本都小于2%,稳定性最优;GBDT在Group2时,稳定性较好,但在Group1、3和4时,*RMSE*变化率的范围为4.2%~5.8%,稳定性稍差;而RF在各情况下,测试集*RMSE*都远高于训练集*RMSE*,*RMSE*变化率最高可达175.7%,存在明显的过拟合现象,稳定性最差。

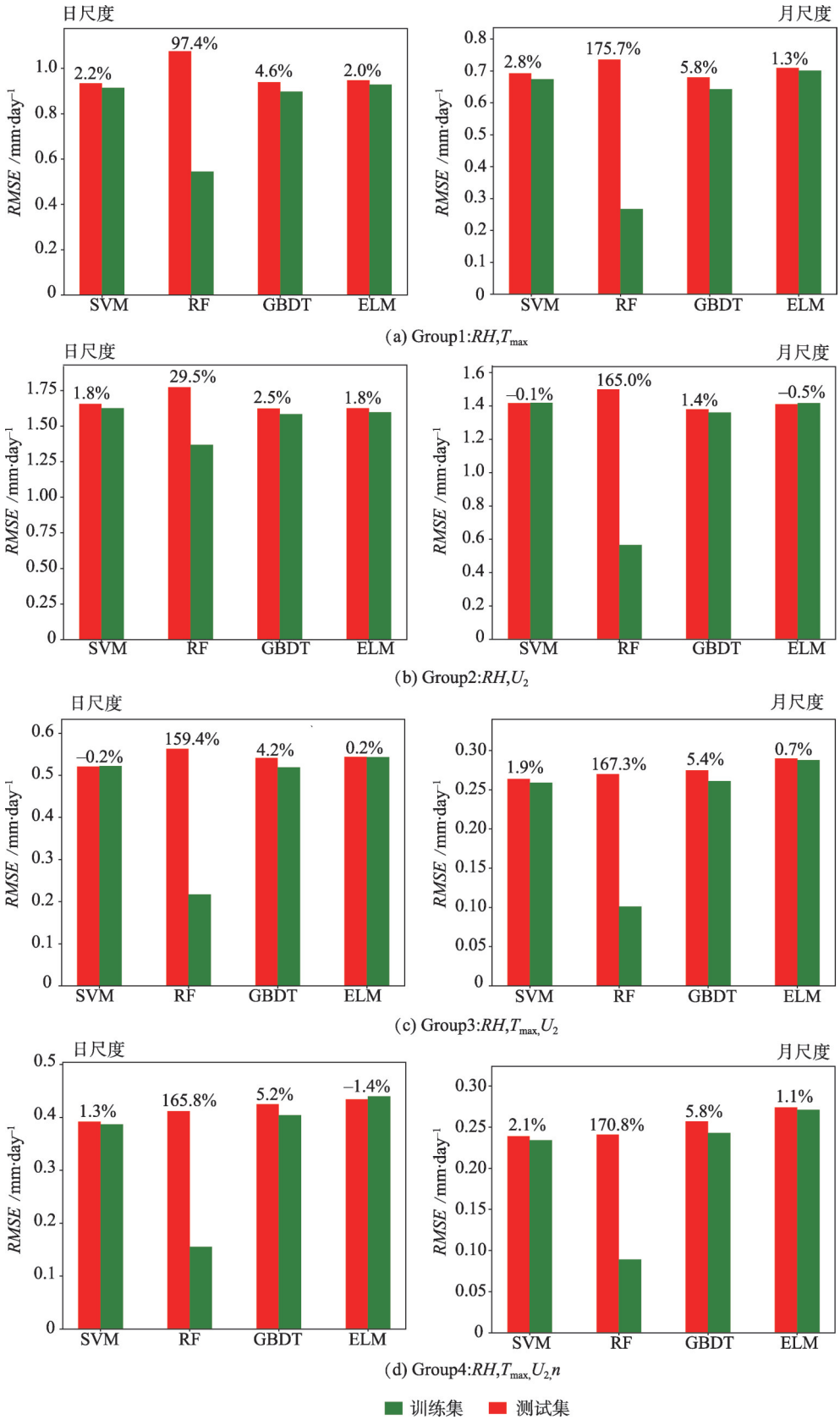
4.4 计算代价分析

各算法在日尺度和月尺度下拟合建模所需时间如图4所示,2种尺度下,GBDT计算代价都是最少,而同为树模型的RF算法计算代价在日尺度平均约为GBDT的4倍,在月尺度上平均约为GBDT的3倍。

ELM和SVM模型计算代价显著高于GBDT和RF,其中SVM计算代价最高。在日尺度,SVM计算代价分别约为ELM、RF和GBDT的1.6、4.5和17.8倍;在月尺度上,分别约为ELM、RF和GBDT的1.7、2.7和8.8倍。

4.5 讨论

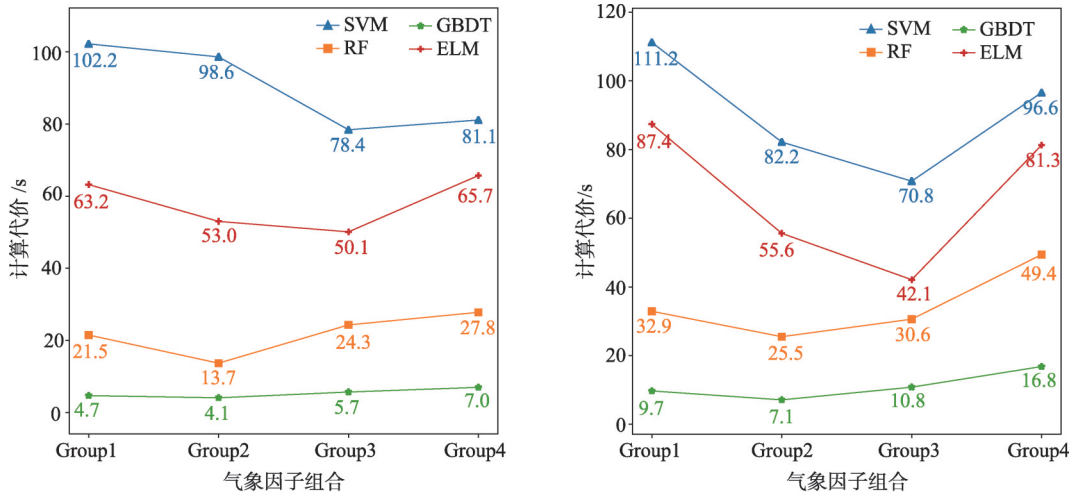
在日尺度上,SVM精度高于GBDT,而在月尺度,GBDT与SVM精度相当。这是由于GBDT是基于树的集成模型,其原理以解决离散的分类问题为目标,导致其在用于连续的回归问题时,与其他模型相比,对数据分布的连续性依赖较高。而本文所使用的月尺度数据是由日尺度数据平均得到,与日尺度数据相比,数据的特征分布被缩小,因此GBDT更容易学习到精度高的模型。GBDT在两种时间尺度上,精度都高于RF,这主要是因为GBDT在训练过程中,后一颗树会对前一颗树的较差结果进行权值校正,逐步提升拟合结果;而平均方法的RF,无法对精度较差的树进行校正,若精度较差的树的数量较多,模型最终精度将被降低。ELM算法得到的模型整体上精度比其他算法差,但对比其在测试集和训练集的精度,并没有过拟合现象。Wen等[2]曾



注：百分数表示模型在测试集的RMSE相对在训练集的RMSE的变化率。

图3　日尺度和月尺度下机器学习模型在测试集与训练集的RMSE柱状图

Fig.3　Comparative histogram between average RMSE of machine learning models on test set and training set at daily and monthly scale



(a) 日尺度 (b) 月尺度

图4 日尺度和月尺度下机器学习模型的计算代价折线图

Fig.4 Line charts of machine learning models' computational costs at daily scale and monthly scale

在干旱区对比了SVM和ANN对日ET_0的拟合效果，发现ANN模型精度低于SVM模型，与本文结果一致。因此，ELM模型较差的主要原因在于其是一种单隐藏层ANN模型，与其他模型相比，对干旱区域蒸散的复杂过程拟合效果较差。

在稳定性方面，SVM、ELM稳定性较高，GBDT次之，而RF算法存在严重的过拟合现象，这与Hassan等[25]在使用SVM和RF拟合太阳辐射时得到的结果一致，同样也是因为树模型对数据分布的连续性依赖较高，当训练集没有包含测试集的数据分布时，其在测试集的精度将会降低。

在计算代价方面，SVM和ELM在日尺度和月尺度的计算代价都高于GBDT和RF，与Fan等[7]的研究结果相一致。一方面，主要因为SVM和ELM算法所需搜索的超参数比GBDT和RF多，而每增加一个超参数，计算代价将成倍增长；另一方面，SVM和ELM数学原理与其他算法相比更为复杂，目标函数求解更为耗时。

在日尺度上，算法适用性由高到低为SVM、GBDT、ELM和RF；在月尺度上，算法适用性由高到低分别为GBDT、SVM、RF和ELM。综合考虑，SVM和GBDT在2种尺度下都有较高的算法，可代替FAO-56 PM公式在新疆区域范围内对ET_0进行估算。

5 结语

本文基于中国气象数据网提供的新疆地区40年（1980—2019年）的地面气候资料日值数据集，在日和月尺度下，通过对气象因子进行敏感性分析，形成不同的气象因子组合；然后使用4种机器学习算法SVM、RF、GBDT和ELM，以FAO-56 PM计算值为标准值，对新疆地区的ET_0进行了拟合建模；最后，从拟合精度、稳定性和计算代价3个方面对模型进行评价，研究表明：

（1）在新疆地区，RH、T_{max}、U_2、n、T_{min}和T_{avg} 6项气象因子中，ET_0对RH、T_{max}和U_2敏感系数级别为高，2种尺度下，平均敏感系数分别为−0.516、0.283和0.266；n为中等，平均敏感系数为0.124；T_{min}和T_{avg}为低，平均敏感系数分别为−0.016和−0.003。

（2）在日尺度，各算法在RH、T_{max}、U_2和n 4项气象因子为输入时精度较高（$RMSE$<0.5 mm/day，R^2>0.95）；在月尺度，各算法使用RH、T_{max}和U_2 3项气象因子便可对ET_0进行精确估算。SVM和GBDT 2种算法在日尺度和月尺度都有较好的适用性，可在相应尺度下使用较少气象因子替代FAO-56 PM公式对ET_0进行估算。

本研究对于新疆地区不同时间尺度下，使用较少气象因子估算参考作物蒸散量具有重要意义，研究结果可用于新疆地区及其他气候类似区域的水资源管理和分配、气候变化研究等有关领域。

参考文献(References)：

[1] Allen R G, Pereira L S, Raes D, et al. Crop evapotranspiration- guidelines for computing crop water requirements [M]. Rome: Food and Agriculture Organization of the United Nations, 1998.

[2] Wen X, Si J, He Z, et al. Support-vector-machine-based

models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions[J]. Water Resources Management, 2015,29(9):3195-3209.

[3] Wu L, Fan J. Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration[J]. PLoS One, 2019,14(5):e0217520.

[4] 王升,付智勇,陈洪松,等.基于随机森林算法的参考作物蒸发蒸腾量模拟计算[J].农业机械学报,2017,48(3):302-309. [Wang S, Fu Z Y, Chen H S, et al. Simulation of reference evapotranspiration based on random forest method [J]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(3):302-309.]

[5] Huang G, Wu L, Ma X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions[J]. Journal of Hydrology, 2019,574:1029-1041.

[6] Fan J, Ma X, Wu L, et al. Light gradient boosting machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data[J]. Agricultural Water Management, 2019,225:105758.

[7] Fan J, Yue W, Wu L, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China[J]. Agricultural and Forest Meteorology, 2018,263:225-241.

[8] 赵文刚,马孝义,刘晓群,等.基于神经网络算法的广东省典型代表站点ET_0简化计算模型研究[J].灌溉排水学报,2019,38(5):91-99. [Zhao W G, Ma X Y, Liu X Q, et al. Using neural network model to simplify ET_0 calculation for representative stations in Guangdong province[J]. Journal of Irrigation and Drainage, 2019,38(5):91-99.]

[9] Feng Y, Peng Y, Cui N, et al. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data[J]. Computers and Electronics in Agriculture, 2017,136:71-78.

[10] Ferreira L B, Da Cunha F F, De Oliveira R A, et al. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM: A new approach[J]. Journal of Hydrology, 2019,572:556-570.

[11] 张皓杰,崔宁博,徐颖,等.基于ELM的西北旱区参考作物蒸散量预报模型[J].排灌机械工程学报,2018,36(8):779-784. [Zhang H J, Cui N B, Xu Y, et al. Prediction for reference crop evapotranspiration in arid northwest China based on ELM[J]. Journal of Drainage and Irrigation Machinery Engineering, 2018,36(8):779-784.]

[12] 吴立峰,鲁向晖,刘小强,等.蝙蝠算法优化极限学习机模拟参考作物蒸散量[J].排灌机械工程学报,2018,36(9):802-805,829. [Wu L F, Lu X H, Liu X Q, et al. Simulation of reference crop evapotranspiration by using bat algorithm optimization based extreme learning machine[J]. Journal of Drainage and Irrigation Machinery Engineering, 2018,36(9):802-805,829.]

[13] 冯禹,崔宁博,龚道枝,等.基于极限学习机的参考作物蒸散量预测模型[J].农业工程学报,2015,31(S1):153-160. [Feng Y, Cui N, Gong D, et al. Prediction model of reference crop evapotranspiration based on extreme learning machine[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015,31(S1):153-160.]

[14] Wu L, Zhou H, Ma X, et al. Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio- inspired optimization algorithms: application in contrasting climates of China[J]. Journal of Hydrology, 2019,577:123960.

[15] Reis M M, Da Silva A J, Zullo Junior J, et al. Empirical and learning machine approaches to estimating reference evapotranspiration based on temperature data[J]. Computers and Electronics in Agriculture, 2019,165:104937.

[16] 刘悦,崔宁博,李果,等.近56年西南地区四季参考作物蒸散量变化成因分析[J].节水灌溉,2018(12):54-59. [Liu Y, Cui N B, Li G, et al. Attribution analysis of seasonal reference crop evapotranspiration in southwest China in recent 56 years[J]. Water Saving Irrigation, 2018(12):54-59.]

[17] Lenhart T, Eckhardt K, Fohrer N, et al. Comparison of two different approaches of sensitivity analysis[J]. Physics and Chemistry of the Earth, 2002,27(9-10):645-654.

[18] 周志华.机器学习[M].北京:清华大学出版社,2016:133-139. [Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016:133-139.]

[19] Kisi O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree[J]. Journal of Hydrology, 2015,528:312-320.

[20] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.

[21] 李航.统计学习方法[M].北京:清华大学出版社,2012:55-75. [Li H. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012:55-75.]

[22] 张雷,王琳琳,张旭东,等.随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J].生态学报,2014,34(3):650-659. [Zhang L, Wang L L, Zhang X D, et al. The basic principle of random forest and its applications in ecology: A case study of Pinus yunnanensis[J]. Acta Ecologica Sinica, 2014,34(3):650-659.]

[23] Friedman J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002,38(4):367-378.

[24] Huang G B, Zhu Q Y, Siew C K, et al. Extreme learning machine: a new learning scheme of feedforward neural networks. Proceedings of the 2004 Ieee International Joint Conference on Neural Networks[C]. New York: IEEE, 2004:985-990.

[25] Hassan M A, Khalil A, Kaseb S, et al. Exploring the potential of tree-based ensemble methods in solar radiation modeling[J]. Applied Energy, 2017,203:897-916.