基于核梯度提升树的森林高度估测方法

李建更[1], 张 尹[1], 刘迎春[2]
(1. 北京工业大学信息学部, 北京 100124; 2. 国家林业和草原局调查规划设计院, 北京 100714)

摘 要: 针对大光斑激光雷达波形数据扰动大、树高分布不均匀的问题,基于 Boosting 集成算法的思想, 提出了一种改进的核函数——核梯度提升树(kernel gradient boosting decision tree, KeGBDT). KeGBDT 通过梯度提升树叶子节点的输出值计算连接函数的权值, 使用连接函数的加权作为核函数的表达形式, 从而避免叶子节点中观测值分布不均匀造成的误差. 在实验部分,使用星载激光雷达(geoscience laser altimeter system, GLAS)数据提取的波形特征作为森林高度估测数据集, 在该数据集上将 KeGBDT 与核随机森林(kernel random forests, KeRF)、线性核、高斯核等常用核函数在岭回归和支持向量回归(support vector regression, SVR)算法中进行了森林高度估测对比实验. 另外,基于 KeGBDT 的岭回归和 SVR 模型与线性回归、梯度提升树(gradient boosting decision tree, GBDT)、随机森林等回归算法进行了森林高度估测对比分析. 实验结果表明, 基于 KeGBDT 的回归算法在决定系数与均方根误差两方面都优于常用核函数与回归算法,可以有效减小森林高度估测模型的回归误差.

关键词: 核梯度提升树; 核函数方法; 激光雷达; 森林高度; 机器学习; 特征提取
中图分类号: TP 79 **文献标志码**: A **文章编号**: 0254 - 0037(2021)10 - 1113 - 09
doi: 10.11936/bjutxb2019100013

Forest Height Estimation Method Based on Kernel Gradient Boosting Decision Tree

LI Jiangeng[1], ZHANG Yin[1], LIU Yingchun[2]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Academy of Inventory and Planning, National Forestry and Grassland Administration, Beijing 100714, China)

Abstract: To solve the problem of the large data disturbance and high variance of spatial distribution of trees height in the waveform of large-footprint light detection and ranging (LiDAR), a kernel function, called kernel gradient boosting decision tree (KeGBDT) was introduced in this paper. The weight of the connection function through the output value of the leaf node in decision tree was calculated by KeGBDT, and the weighted sum of the connection function was used as the expression of the kernel function. Therefore, the error caused by the uneven distribution of observation values in the leaf nodes was avoided. In the experimental part, the waveform feature from the geoscience laser altimeter system (GLAS) data was used as the forest height estimation dataset. The KeGBDT was compared with kernel random forests (KeRF), linear kernel, Gaussian kernel and other common kernel functions, and the ridge regression and support vector regression were compared based on KeGBDT with the other regression algorithms, such like linear regression, GBDT and random forests, in forest height estimation task. Results show that the regression algorithm based on KeGBDT is superior to the commonly used kernel function and regression algorithm in both *R*-squared and root mean square error, and KeGBDT can

收稿日期: 2019-10-21
基金项目: 国家自然科学基金资助项目(61873008); 北京市自然科学基金资助项目(4182008)
作者简介: 李建更(1965—), 男, 教授, 主要从事模式识别及其应用方面的研究, E-mail: lijg@bjut.edu.cn

effectively reduce the bias of the estimation model of forest height.

Key words: kernel gradient boosting decision tree (KeGBDT); kernel function method; light detection and ranging (LiDAR); forest height; machine learning; feature extraction

森林是陆地生态系统的重要组成部分，在调节全球气候、维持全球碳平衡、减缓温室气体质量浓度增长等方面起着至关重要的作用[1]. 森林高度是评估森林性能和状态、研究全球碳循环和碳汇的重要数据[2]. 激光雷达(light detection and ranging, LiDAR)是目前估测森林高度最精确的遥感方法，具有实时性强、采集范围广、相对成本低等优点，根据光斑直径可分为小光斑 LiDAR[3] 和大光斑 LiDAR[4]. 小光斑 LiDAR 主要用于采集高分辨率的激光点云数据，而大光斑 LiDAR 通常用于全波形数据的采集. 由于成本问题，全球尺度的森林高度主要选用卫星搭载的大光斑 LiDAR 系统进行估测.

在使用大光斑 LiDAR 数据进行森林高度估测时，其主要思路为对全波形数据进行特征提取，随后使用获取的特征通过机器学习方法反演森林高度[4]. 森林高度反演模型中常用的机器学习算法包括多元线性回归、指数回归、对数回归、幂函数回归、支持向量机、随机森林等[5-9]，但在实际的森林高度估测中，误差来源多样且难以避免，例如：光斑位置解算造成的位置偏差以及云层等影响导致的波形失真等. 支持向量机具有坚实的理论基础、良好的鲁棒性、泛化能力强等优点[10]，被广泛用于基于大光斑 LiDAR 的森林高度估测. 2018 年，Lee 等[5] 使用基于高斯核的支持向量回归(support vector regression, SVR)算法估测机载大光斑 LiDAR 的森林高度，获得了较好的实验效果. 由于随机森林(random forests, RF)算法的随机子空间在样本筛选的过程中可以有效地降低异常点带来的干扰，Pourrahmati 等[6] 利用随机森林算法进行大光斑数据的森林高度估测，模型回归性能良好.

光斑照射区域的地表特征会导致大光斑 LiDAR 的波形数据产生畸变，从而影响高度估测模型的准确性，因此，需要对波形特征进行处理从而增强森林高度估测模型对不同地形的适应性. Sun 等[11] 利用高斯分解法将波形转化成累积量计算波形高度分位数，在地形平坦的地区具有较好的估测性能. Lefsky 等[12] 提出波形参数法，在波形高度分位数的基础上加入波形前沿长度和波形后沿长度等受地形影响较大的波形特征，但森林高度估测模型的准确性并没有显著的提升. 胡凯龙等[13] 提出了地形辅助法，利用坡度信息校准地形带来的偏差值以减小模型的估测误差，提升了模型的地形适应性. 汪垚等[14] 提出了模型辅助法，使用地形坡度模拟裸地回波校准地形坡度对波形产生的展宽. 该方法具有较强的地形适应性，但仍未完全解决复杂林况数据的误差问题.

核函数作为输入空间到特征空间的映射可以使特征空间的性质产生改变，从而降低人工提取特征对数据描述的误差并提高预测模型的准确性，常被应用于基于全波形数据的森林高度估测[5]. 但传统核函数方法的性能对参数选择与优化方法的依赖性较强. 与核函数方法相比，随机森林具有训练参数少、对超参数不敏感等优势. Scornet[15] 利用随机森林与核函数之间的相关性，通过理论推导得出基于随机森林的核函数——核随机森林(kernel random forests, KeRF). KeRF 赋予每一个观测值相同的权重，避免了由于叶子节点的观测值非均匀分布导致的回归误差.

为解决大光斑波形扰动大、森林高度分布不均匀的问题，本文提出了一种基于核梯度提升树的森林高度估测方法，使用星载激光雷达(geoscience laser altimeter system, GLAS)的大光斑波形数据进行森林高度估测. 本文所提出的方法使用梯度提升树(gradient boosting decision tree, GBDT)叶子节点连接函数的权重构建新的核函数——核梯度提升树(kernel gradient boosting decision tree, KeGBDT). KeGBDT 赋予每个观测值相应的权重，有效避免叶子节点中观测值分布不均匀造成的误差. 对于波形数据，本文利用地形信息模拟裸地回波计算分位数特征，并使用主成分分析(principal component analysis, PCA)提取反映森林垂直结构的波形分布特征作为森林高度估测数据集. 在实验部分，将所提出的 KeGBDT 嵌入岭回归和 SVR 算法，并与常用的森林高度估测算法进行实验结果对比.

1 研究数据

1.1 GLAS 数据与研究区域

2003 年，美国国家宇航局发射了搭载着 GLAS 的激光测高卫星. GLAS 于 2003—2009 年的回传数

据被用于全球尺度的植被高度和生物量估测. GLAS 共有 3 个激光器，激光器设置的发射脉冲宽度为 4 ns，大光斑采样间隔为 170 m，光斑直径为 70 m[8]. 本研究采用的 GLAS 数据来自美国国家航空航天局戈达德太空飞行中心（https://nsidc.org），主要为 GLA01 的全波形数据和 GLA14 中光斑点的经度、纬度、高程等信息，获取时间为 2006 年 6 月 8 日.

研究区位于加拿大和美国阿拉斯加州的北方森林地区，地处 50.0°N—70.0°N，90°W—160.0°W，具有广袤的森林. 其地貌多样，包含森林、草原、山地、荒地、湿地、湖泊，同时富含森林资源、生物多样性和矿产等重要资源. 森林地区主要为云杉、落叶松、冷杉、松树等针叶树.

研究所需的地形坡度和森林高度数据，通过 LiDAR360、Arcgis 软件对机载 LiDAR 采集的激光点云数据进行处理可得. 飞机搭载着由 Optech 公司生产的 ALTM3100 系统，以约 1 850 m 的飞行高度、270 km/h 的飞行速度、50 kHz 的发射脉冲频率以及 0.7 m^2 的点云密度飞过本文使用 GLAS 数据的覆盖区域并获取与其飞行轨迹重叠的点云数据.

1.2　数据预处理与特征提取

GLA01 全波形数据记录了每纳秒波形强度经归一化处理后的数值，每个大光斑点记录的波长为 544 ns. 大光斑波形实质上是由各物体对发射激光脉冲的反射回波叠加而成. 大气干扰造成的回波被称为背景噪声，其几乎存在于每一个大光斑波形中，因此，在波形特征提取前需将其滤去. 首先，评估背景噪声，选取大光斑 LiDAR 原始回波波形前后 30 ns 的背景噪声部分，分别计算出均值 μ_1 和 μ_2，选用 μ_1、μ_2 中较小的一个值 $\mu_{\min}$，并计算其所对应的标准差 σ，设 $\mu_{\min}+3\sigma$ 为背景噪声. 因背景噪声存在于整体波形中，故对波形数据统一减去背景噪声. 随后，采用高斯滤波对波形进行平滑处理. 为确保光斑点对应的波形为森林回波，筛选光斑位置存在对应森林高度值的波形数据.

森林回波主要由地面回波和林木回波组成. 地面回波受坡度影响，随坡度的增大产生展宽，易导致森林高度的估测值偏大，甚至失真. 为弥补地形坡度造成的误差，本文使用基于模拟裸地回波的高度分位数. 假设裸地回波为高斯分布，然后利用地形坡度、发射脉冲波长、大光斑半径以及高斯公式计算模拟裸地回波的公式，即

$$y=a\exp\left(\frac{-(x-\mu)^2}{2(\sigma+\beta r\tan\theta)^2}\right) \tag{1}$$

$$\sigma=\sqrt{-(cw)^2/(2\lg 0.5)} \tag{2}$$

式中：$a=1$；$\mu=0$；β 为转换系数，取 0.5；r 为光斑半径；θ 为地形坡度；c 为光速；w 为发射脉冲波长.

模拟裸地回波结束位置与大光斑波形的结束位置相对应，将大光斑滤波后波形能量的 50%、75% 的位置 RH_{50} 和 RH_{75} 与模拟裸地回波波形能量 50%、75% 的位置 Hg_{50} 和 Hg_{75} 作差，得到基于模拟裸地回波的高度分位数 RH_{C50} 和 RH_{C75}，公式为

$$\frac{1}{2}\sum_{i=1}^{n}f(i)=1-\sum_{i=1}^{x_1}f(i) \tag{3}$$

$$\frac{1}{4}\sum_{i=1}^{n}f(i)=1-\sum_{i=1}^{x_2}f(i) \tag{4}$$

$$\frac{1}{2}\sum_{i=1}^{n}g(i)=1-\sum_{i=1}^{x_3}g(i) \tag{5}$$

$$\frac{1}{4}\sum_{i=1}^{n}g(i)=1-\sum_{i=1}^{x_4}g(i) \tag{6}$$

$$RH_{C50}=RH_{50}-Hg_{50} \tag{7}$$

$$RH_{C75}=RH_{75}-Hg_{75} \tag{8}$$

式中：$f(i)$ 为大光斑滤波后函数；$g(i)$ 为模拟裸地回波函数；n 为波形长度；x_1、x_2、x_3、x_4 分别为 RH_{50}、RH_{75}、Hg_{50}、Hg_{75}.

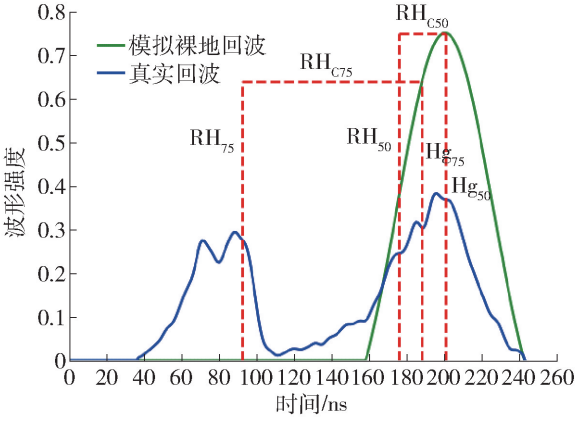

图 1　高度分位数

Fig. 1　Height quantile

高度分位数 RH_{C50} 和 RH_{C75} 矫正了坡度对波形的影响，同时与森林高度具有较强的相关性，但是林木回波反馈的信息不只是森林高度，还包含林木的垂直结构、冠层分布. 因此，仅通过高度分位数反演森林高度具有一定局限性. 除校正后的高度分位数外，本文对原始波形的强度进行了主成分分析，提取了共计六维的主成分特征. 基于模拟裸地回波的高度分位数和主成分特征为本文使用的森林高度估测数据集，数据集共计 6 条大光斑条带，其中训练集包含 8 000 个光斑点数据，验证集包含

2 000 个光斑点数据.

表 1 森林高度估测数据集

Table 1 Dataset of forest height estimation

特征分类	波形特征
高度分位数	RH_{C10}、RH_{C20}、RH_{C30}、RH_{C40}、RH_{C50}、RH_{C60}、RH_{C70}、RH_{C75}、RH_{C80}、RH_{C90}、RH_{C100}
主成分特征	P_1、P_2、P_3、P_4、P_5、P_6

2 方法

2.1 梯度提升树

梯度提升算法(gradient boosting machine, GBM)是一种将多个弱学习器加权组合为强学习器的加法模型, 通过最小化损失函数得到当前迭代步骤的模型伪残差及最优解方向, 在处理非正态分布的样本时能够更好地拟合离散点. 其中, 伪残差为当前迭代步骤中令损失函数最小化的每一个数据点处的模型梯度. 具体地,对于一个由$x\in\mathbb{R}^d$输入与$y\in\mathbb{R}$输出组成的回归问题$y=F(x)$,给定训练数据集$D_{\text{train}}:=\{(x_1,y_1),\cdots,(x_N,y_N)\}$与损失函数$L(y,F(x))$, 梯度提升算法模型可表示为

$$F(x)=\alpha\sum_{m=1}^{M}f_m(x) \tag{9}$$

式中: M 为迭代次数;$f_m(x)$为第 m 次迭代的弱学习器;α 为学习率. $f_m(x)$通过优化目标函数$\min\sum_{i=1}^{N}L(y_i,F_{m-1}(x_i)+f_m(x_i))$得到, 对于回归问题, 损失函数可以使用最小二乘损失(least-squares, LS)、最小绝对偏差损失(least absolute deviation, LAD)或 Huber 损失等方法进行计算.

GBDT 是一种使用决策树作为弱学习器的梯度提升算法, 在每一次迭代步骤中使用前次迭代模型的伪残差训练一棵新决策树以顺序生成弱学习器. 在完成算法迭代训练后, GBDT 模型可表示为

$$F(x)=\alpha\sum_{m=1}^{M}f_m(x)=\alpha\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}1_{x\in R_{mj}} \tag{10}$$

式中:J 为第 m 次迭代的回归树中叶子节点个数;f_m为 GBDT 中单棵决策树的模型;c_{mj}为回归树中叶子节点的最优输出值;$R_{mj}(j=1,2,\cdots,J)$为叶子节点所属特征空间. 当 x 与 R_{mj}属于同一特征子空间, $1_{x\in R_{mj}}=1$.

对梯度提升树模型进行初始化,令$f_0(x)=0$, 在第 m 个迭代步骤中, 首先,基于优化目标求取当前迭代步骤每一个数据点对应的伪残差 r_{mi}. 在本文中,使用最小二乘损失 $L(y,F(x))=\|y-F(x)\|^2/2$,其优化目标遵循

$$J(x)=\min\frac{1}{2}\sum_{i=1}^{N}\|y_i-(F_{m-1}(x_i)+\alpha f_m(x_i))\|^2 \tag{11}$$

式中 r_{mi}为损失函数在数据点 x_i 处模型的负梯度. 当模型输入为 x_i 时, 求取损失函数对模型输出$F_{m-1}(x_i)$的偏导,可得

$$r_{mi}=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}=-\frac{\partial\|y-F_{m-1}(x_i)\|^2/2}{\partial F_{m-1}(x_i)}=y_i-F_{m-1}(x_i)=y_i-\sum_{p=1}^{m-1}f_p(x_i) \tag{12}$$

对训练集中的每一个数据点求取伪残差 r_{mi}, 对于当前迭代步骤,可以使用伪残差生成一组新训练数据 $D_m:=\{(x_1,r_{m1}),\cdots,(x_N,r_{mN})\}$. 在 D_m 上训练一棵新的回归树 $f_m(x)$, 则回归树 $f_m(x)$叶子节点的最优值为

$$c_{mj}=\arg\min_{c}\sum_{x_i\in R_{mj}}L(y_i,F_{m-1}(x_i)+c)=\frac{\sum_{i=1}^{N}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in R_{mj}}}{\sum_{i=1}^{N}1_{x_i\in R_{mj}}} \tag{13}$$

对当前迭代步骤的强学习器进行更新可得

$$F_m(x)=F_{m-1}(x)+\alpha f_m(x) \tag{14}$$

综上所述, 在算法完成迭代更新后, 带入决策树模型, 获得的强学习器 $F_M(x)$可表示为

$$F_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}\frac{\sum_{i=1}^{N}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in R_{mj}}}{\sum_{i=1}^{N}1_{x_i\in R_{mj}}}1_{x\in R_{mj}}=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}1_{x\in R_{mj}} \tag{15}$$

2.2 核梯度提升树 KeGBDT

KeRF 是一种更具解释性的随机森林模型, 其使用随机森林中样本间的连接函数作为核函数. KeRF 可直接用于岭回归、支持向量机等核方法,相比于常用核函数具有更好的核函数性能[16]. 给定训练数据集 $D_{\text{train}}:=\{(x_1,y_1),\cdots,(x_N,y_N)\}$,KeRF 的预测模型可表示为

$$\widetilde{K}_{M,n}(x,x_i)=\frac{1}{M}\sum_{m=1}^{M}1_{x_i\in A_n(x,\Theta_m)} \tag{16}$$

$$m_{M,n}(x,\Theta_1,\cdots,\Theta_M)=\frac{1}{M}\frac{\sum_{i=1}^{N}y_i\widetilde{K}_{M,n}(x,x_i)}{\sum_{l=1}^{N}\widetilde{K}_{M,n}(x,x_l)} \tag{17}$$

式中：M 为随机森林中决策树的个数；Θ_m 为第 m 次迭代时决策树的分裂模式；$m_{M,n}(x,\Theta_1,\cdots,\Theta_M)$ 为随机森林回归函数. 定义 $A_n(x,\Theta_m)$ 为森林中包含 x 的节点，由 Θ_m 和数据集 D_{train} 决定，当 x_i 与 x 在决策树模型中连接至同一叶子节点时，$1_{x_i\in A_n(x,\Theta_m)}$ 取值为 1[15]. 实际上，对于固定的训练数据集 D_{train}，核函数 $\widetilde{K}_{M,n}$ 可表示输入 x 在森林中与 x_i 连接至同一叶子节点的后验概率. Arlot 等[17]于 2014 年的研究表明，特殊的森林模型同样可以推导为森林核估计的形式，并与原森林模型表现出一致性. 对于回归任务，GBDT 算法通过线性搜索迭代减小损失函数以获得强学习器，相比于随机森林方法具有更小的统计偏差. 因此，本文基于 GBDT 方法提出了一种改进的核函数 KeGBDT.

在 GBDT 方法中，对于第 m 个迭代步骤生成的决策树 $f_m(x,\Theta_m)$，将式(13)代入决策树可得

$$f_m(x,\Theta_m)=\sum_{j=1}^{J}\frac{\sum_{i=1}^{N}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in R_{mj}}}{\sum_{i=1}^{N}1_{x_i\in R_{mj}}}1_{x\in R_{mj}}=\frac{\sum_{i=1}^{N}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)\sum_{j=1}^{J}1_{x_i\in R_{mj}}1_{x\in R_{mj}}}{\sum_{i=1}^{N}\sum_{j=1}^{J}1_{x_i\in R_{mj}}} \tag{18}$$

为表示输入 x 与训练数据集间的相关性，将 $A_n(x,\Theta_m)$ 代入式(18)，则 $f_m(x,\Theta_m)$ 可表示为

$$f_m(x,\Theta_m)=\frac{\sum_{i=1}^{N}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in A_n(x,\Theta_m)}}{\sum_{i=1}^{N}1_{x_i\in A_n(x,\Theta_m)}} \tag{19}$$

将单棵决策树模型 $f_m(x,\Theta_m)$ 代入 GBDT 模型，得到

$$F_{M,n}(x,\Theta_1,\cdots,\Theta_M)=\sum_{m=1}^{M}\sum_{i=1}^{N}\frac{\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in A_n(x,\Theta_m)}}{N_n(x,\Theta_m)} \tag{20}$$

将 GBDT 模型视为输入 x 与数据点 (x_i,y_i) 在每一颗决策树 $f_m(x,\Theta_m)$ 中的连接函数，对于每一个数据点 (x_i,y_i)，对其连接函数赋予权重

$$w_{i,m}=\frac{\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)}{y_i}1_{x_i\in A_n(x,\Theta_m)} \tag{21}$$

则 GBDT 模型可表示为

$$F_{M,n}(x,\Theta_1,\cdots,\Theta_M)=\sum_{m=1}^{M}\sum_{i=1}^{N}\frac{y_i\frac{1}{y_i}\left(y_i-\sum_{p=1}^{m-1}r_{pi}\right)1_{x_i\in A_n(x,\Theta_m)}}{N_n(x,\Theta_m)}=\frac{1}{\sum_{m=1}^{M}N_n(x,\Theta_m)}\sum_{m=1}^{M}\sum_{i=1}^{N}y_iw_{i,m}=\frac{\sum_{i=1}^{N}y_i\sum_{m=1}^{M}w_{i,m}}{\sum_{m=1}^{M}N_n(x,\Theta_m)} \tag{22}$$

$F_{M,n}(x,\Theta_1,\cdots,\Theta_M)$ 等同于梯度提升树中所有包含 x 的叶子节点对应观测值的加和，每一个观测值都通过其在森林中出现的次数进行加权. 因此，空节点不参与预测，有效地避免了因空节点造成的回归误差. 在不使用叶子节点观测值的权重时，无权重 KeGBDT 可以表示为连接函数和的形式

$$K'_{M,n}(x,x_i)=\sum_{m=1}^{M}1_{x_i\in A_n(x,\Theta_m)} \tag{23}$$

$\widetilde{K}_{M,n}(x,x_i)$ 为 $m_{M,n}(x,\Theta_1,\cdots,\Theta_M)$ 中 y_i 被赋予的权重，即 x_i 与 x 在同一叶子节点中出现的概率. $\sum_{m=1}^{M}w_{i,m}$ 为 $F_{M,n}(x,\Theta_1,\cdots,\Theta_M)$ 中 y_i 被赋予的权重，即 x_i 与 x 在同一叶子节点时残差占 y_i 的比率. 与 $m_{M,n}(x,\theta_1,\cdots,\theta_M)$ 相比，$F_{M,n}(x,\theta_1,\cdots,\theta_M)$ 中 y_i 的权重分配方式对残差较大的数据点分配更多的权重，可以加强模型对离散点的学习. 因此，本文结合核函数思想，利用不同样本在 GBDT 算法叶子节点样本空间中的分布信息，根据 $\sum_{m=1}^{M}w_{i,m}$ 计算得出

$$K_{M,n}(x,x_i)=\sum_{m=1}^{M}w_{i,m}=\sum_{m=1}^{M}\frac{y_i-\sum_{p=1}^{m-1}r_{pi}}{y_i}1_{x_i\in A_n(x,\Theta_m)} \tag{24}$$

算法 1　核梯度提升树.

初始化：训练数据集 $D_{\text{train}}=\{(x_1,y_1),\cdots,(x_N,y_N)\}$，迭代次数 M，学习率 α，损失函数 $L=0.5\times\|y-F(x)\|^2$，$F(x)=0$.

1. for $m=1$ to M do
2. for $i=1$ to N do

3. $r_{mi} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} = y_i - F(x_i)$

4. end for

5. 使用$(X_i, r_{mi})(i=1, 2, \cdots, N)$构建一个新的分类回归树(classification and regression tree, CART)的分裂变量.

6. $F_m = F_{m-1} + \alpha f_m$

7. for $i=1$ to N do

8. $w_{i,m} = \frac{\left(y_i - \sum_{p=1}^{m-1} r_{pi}\right)}{y_i} 1_{x_i \in A_n(x, \Theta_m)}$

9. end for

10. end for

11. 计算核梯度提升树:$K_{M,n}(x, x_i) = \sum_{m=1}^{M} w_{i,m}$

3 实验结果与分析

3.1 实验设计

为证明 KeGBDT 相比传统的核函数可以更有效地计算森林高度估测数据集在高维空间投影的内积,以及基于 KeGBDT 的核方法相比其他机器学习算法能够减小森林高度估测值的误差,本文共设计了 3 个实验内容,每组实验都在森林高度估测数据集上进行. 其中,实验 1 通过判断 KeGBDT 在训练数据集上核函数矩阵的半正定性,验证核函数的有效性;实验 2 通过对比基于 KeGBDT 与其他核函数的核方法的森林高度估测结果,以及基于 KeGBDT 的核方法与常用回归算法的森林高度估测结果,证明了基于 KeGBDT 的森林高度估测方法的有效性;实验 3 通过对比决策树数量对基于 KeGBDT 与 KeRF 的 SVR 算法的影响,证明 KeGBDT 所需决策树数量更少,计算量更小.

在实验 1 与实验 2 中,KeGBDT 的具体参数如下:决策树个数为 25,学习率为 0.1,损失函数使用最小二乘损失,单棵决策树的最大深度为 3,叶子节点数不受限制. 在实验中,通过决定系数(coefficient of determination, R^2)和均方根误差(root mean squared error, RMSE)2 个指标评价模型的森林高度估测效果. R^2和 RMSE 公式分别为

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - y_i^*)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{25}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - y_i^*)^2} \tag{26}$$

式中:y_i 为实际森林高度;y_i^* 为估测森林高度;$\bar{y}$ 为实际森林高度的平均值.

3.2 核函数有效性判定与算法复杂度分析

本实验为 KeGBDT 的有效性判定实验,对其在训练集上的核函数矩阵的正定性进行分析,通过 Mercer 定理证明核函数的有效性.

在 KeGBDT 中,核函数可以表示为

$$K_{M,n}(x, x_i) = \sum_{m=1}^{M} \frac{y_i - \sum_{p=1}^{m-1} r_{pi}}{y_i} 1_{x_i \in A_n(x, \Theta_m)} \tag{27}$$

即将核函数定义为不同数据点在同一叶子节点中的连接函数的线性组合. 对于 N 组大小的训练数据集,每一棵决策树的核函数可以写为 $N \times N$ 大小的矩阵形式. 基于文献[16]中的证明,单棵决策树的核函数矩阵可以被置换为半正定矩阵的形式. 若单棵决策树的核函数矩阵为 Gram 矩阵,则单棵决策树构成的核函数为有效的半正定核函数,而半正定核函数的线性组合为有效核函数,因此,可以说明 KeGBDT 为有效核函数.

图 2 显示了 KeGBDT 在 4 000 组训练数据上核函数特征矩阵的能量. 为了更直观地展现核函数矩阵的半正定性,对实验中每一个特征值取以 10 为底的对数形式. 从图 2 中可以清晰地看出,主对角线排列的每一个特征值均大于或等于零. 因此,KeGBDT 在训练数据集上的核函数矩阵为 Gram 矩阵. 根据 Mercer 定理,任何半正定的函数均可作为有效的核函数,以此可以证明 KeGBDT 作为核函数的有效性.

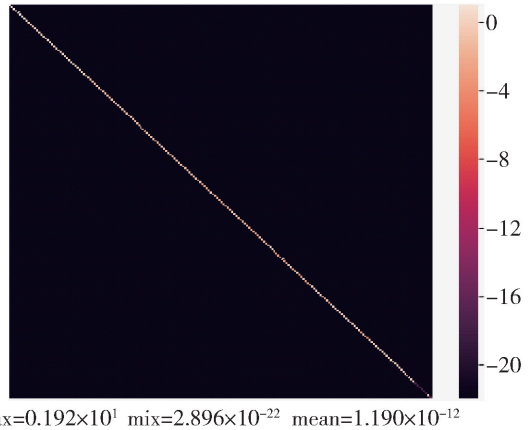

图 2 KeGBDT 核函数矩阵能量

Fig. 2 Kernel function matrix energy of KeGBDT

同时，为了说明算法的计算效率，对 KeGBDT 的算法复杂度进行了分析. 由于本文所提出的 KeGBDT 可以简单且方便地嵌入 SVR 等基于核函数的算法，因此，只针对核函数的计算复杂度进行对比与分析. 在不考虑核方法求取解析解或近似解的复杂度的前提下，给定待测试样本 x，KeGBDT 的时间复杂度为 $O((2N+1)TK)$. 其中：N 为训练集样本个数；T 为决策树个数；K 为决策树叶节点个数. 复杂度包括 $N+1$ 个样本在决策树上的运行时间以及测试样本 x 与训练数据集间叶子节点连接函数的匹配. 若在生成决策树时以 $O(NTK)$ 的空间复杂度存储训练数据在决策树上的叶子节点划分结果，则 KeGBDT 的时间复杂度可以简化至 $O(NTK)$. 可以看出，在完成决策树的构建后，KeGBDT 的计算复杂度与特征维度无关，而是由决策树的规模与数量决定的. KeGBDT 在测试阶段具有训练数据集规模线性范围内的计算复杂度.

3.3　森林高度估测

本实验基于 Sigmoid、线性核、多项式核、高斯核、KeRF、无权重 KeGBDT、KeGBDT，通过岭回归和 SVR 算法进行森林高度估测. 另外，使用线性回归、RF[18]、GBDT、极限梯度提升(extreme gradient boosting，XGBoost)进行了森林高度估测的对比实验. 实验结果如表 2 所示.

如图 3、4 所示，在岭回归和 SVR 算法中，Sigmoid 不适用于实验数据集，无法正确反映训练数据特征的高维分布；使用基于树模型的核函数 KeRF 和 KeGBDT 时，回归模型性能明显提升，可见通过树模型构造的核函数可以有效地处理相关性高的数据集；基于 KeGBDT 的森林高度估测模型明显优于无权重 KeGBDT，说明在核函数中加入权重使实验数据集分布更合理，有效解决了在无权重 KeGBDT 内数据分布集中的问题.

由表 2 所示，在森林高度估测的实验结果中，RF、GBDT、XGBoost 的回归效果较好. 其中，随机森林的回归结果误差最小，说明随机森林行采样和列采样的训练方式可以有效地避免因数据间相关性强导致模型易陷入过拟合的问题. 这一现象在 KeRF 中也有所体现，基于 KeRF 的 SVR 算法的回归结果评价指标 R^2 和 RMSE 都与 RF 相近，但两者的估测结果误差都高于使用 KeGBDT 的岭回归和 SVR 算法. 虽然，由于过拟合问题 GBDT、XGBoost 的回归效果均略逊于 RF，但是基于 KeGBDT 的核方法仅使用较少的决策树棵数就可以完成训练，从而在一定程度上避免了过拟合的问题.

表 2　森林高度估测结果

Table 2　Forest height estimation results

核函数	算法	R^2	RMSE
Sigmoid	岭回归	0.01	4.43
	SVR	−0.08	4.70
线性核	岭回归	0.52	3.04
	SVR	0.65	2.57
多项式核	岭回归	0.67	2.53
	SVR	0.59	2.70
高斯核	岭回归	0.53	3.00
	SVR	0.71	2.46
KeRF	岭回归	0.75	2.24
	SVR	0.78	2.13
无权重 KeGBDT	岭回归	0.62	2.66
	SVR	0.65	2.57
KeGBDT	岭回归	0.81	1.94
	SVR	0.82	1.91
无核	线性回归	0.47	3.19
	RF	0.80	2.01
	GBDT	0.78	2.12
	XGBoost	0.78	2.11

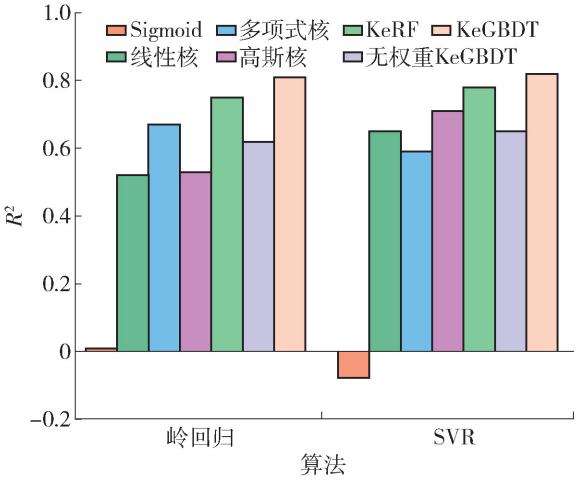

图 3　核函数 R^2 对比

Fig. 3　Comparison of kernel function R-square

图 5 为基于 KeGBDT 的 SVR 算法散点图，回归效果最好. 从图中可见，测试时存在个别与实际大光斑森林高度相差达 10 m 的估测值，为此查看了

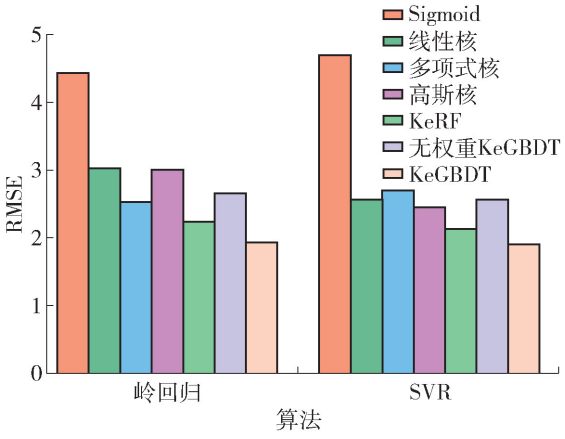

图 4 核函数 RMSE 对比

Fig. 4 Comparison of kernel function RMSE

估测值误差较大的光斑点的对应波形，发现波形与对应的大光斑森林高度不符，应为大光斑位置解算时出现的偏差，导致波形探测位置的森林高度与采集的标签森林高度不符. 但实际森林高度估测过程中，累积误差是难以避免的，森林高度估测模型应具有良好的抗干扰性.

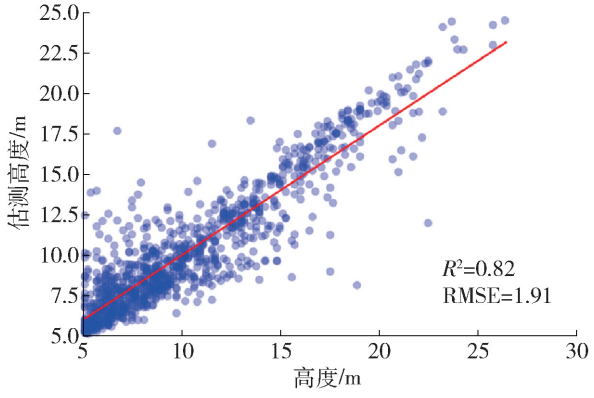

图 5 基于 KeGBDT 的 SVR 算法散点

Fig. 5 Scatter plot of SVR algorithm with KeGBDT

3.4 决策树数量对 KeGBDT 的影响

在本实验中，对比了 SVR 算法分别使用不同决策树数量的 KeGBDT 与 KeRF 时的回归效果.

从图 6、7 中可见，一开始基于 KeGBDT 的 SVR 算法随决策树数量的增大，回归模型的 R^2 提升更快，RMSE 下降更迅速，说明决策树数量较小的情况下更适于选择 KeGBDT；当决策树棵数为 25 时，基于 KeGBDT 的 SVR 算法的回归性能达到最好，而在决策树数量达到 125 棵以上时基于 KeRF 的回归模型性能才达到最好，KeGBDT 所需决策树数量明显小于 KeRF. 因此，为减小决策树数量，缩小计算量，应选用 KeGBDT.

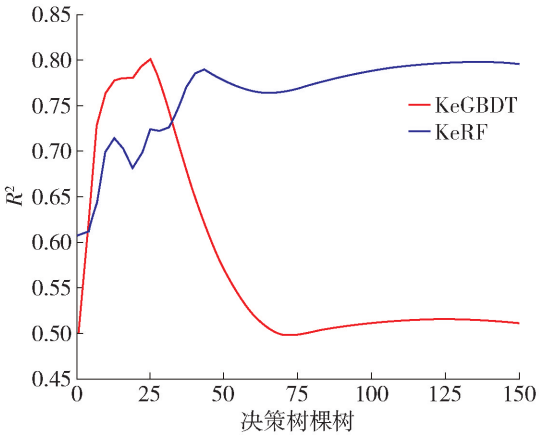

图 6 KeGBDT 与 KeRF 的 R^2 对比

Fig. 6 Comparision of the R-square between KeGBDT and KeRF

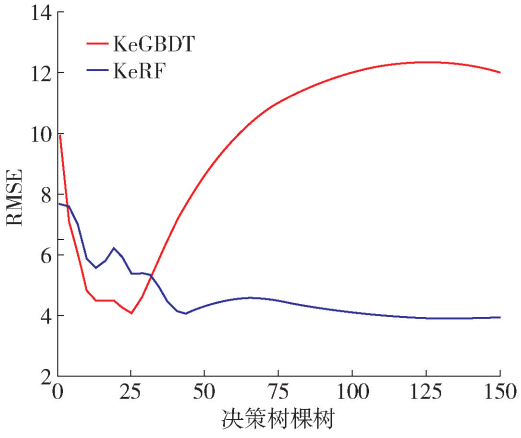

图 7 KeGBDT 与 KeRF 的 RMSE 对比

Fig. 7 Comparision of RMSE between KeGBDT and KeRF

4 结论

1）使用 GBDT 中叶子节点的连接函数的权值构造核函数 KeGBDT，在理论推导和实际构建方面都具有可行性. 本文通过对比实验认为，基于 KeGBDT 的岭回归和 SVR 算法对干扰大、不均匀、相关性强的森林高度估测数据集，回归误差明显小于其他核函数和机器学习算法.

2）相比 KeRF，基于 KeGBDT 的 SVR 算法对于数量较少的决策树能够获得误差最小的回归结果，计算量更小.

3）在森林高度估测过程中，使用基于模拟裸地回波的高度分位数和主成分分析提取的特征向量作为数据集. 在该数据集下，使用基于 KeGBDT 的岭回归和 SVR 算法进行森林高度估测，能够有效减小森林高度估测误差.

参考文献：

[1] WOODWELL G M, WHITAKER R H, REINERS W A, et al. The biota and the world carbon budget[J]. Science, 1978, 199(4325): 141-146.

[2] DRAKE J B, DUBAYAH R O, CLARK D B, et al. Estimation of tropical forest structural characteristics using large-footprint lidar[J]. Remote Sensing of Environment, 2002, 79(2): 305-319.

[3] 马利群, 李爱农. 激光雷达在森林垂直结构参数估算中的应用[J]. 世界林业研究, 2011, 24(1): 41-45.
MA L Q, LI A N. Review of application of LiDAR to estimation of forest vertical structure parameters[J]. World Forestry Research, 2011, 24(1): 41-45. (in Chinese)

[4] 文汉江, 程鹏飞. ICESAT/GLAS 激光测高原理及其应用[J]. 测绘科学, 2005, 30(5): 33-35.
WEN H J, CHENG P F. Introduction to principle of ICESAT/GLAS laser altimetry and its applications[J]. Science of Surveying and Mapping, 2005, 30(5): 33-35. (in Chinese)

[5] LEE J, IM J, KIM K, et al. Machine learning approaches for estimating forest stand height using plot-based observations and airborne LiDAR data[J]. Forests, 2018, 9(5): 268.

[6] POURRAHMATI M R, BAGHDADI N, DARVISHSEFAT A A, et al. Mapping Lorey's height over Hyrcanian forests of Iran using synergy of ICESat/GLAS and optical images[J]. European Journal of Remote Sensing, 2018, 51(1): 100-115.

[7] WANG M, SUN R, XIAO Z. Estimation of forest canopy height and aboveground biomass from spaceborne LiDAR and Landsat imageries in Maryland[J]. Remote Sensing, 2018, 10(2): 344.

[8] POURRAHMATI M R, BAGHDADI N N, DARVISHSEFAT A A, et al. Capability of GLAS/ICESat data to estimate forest canopy height and volume in mountainous forests of Iran[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2017, 8(11): 5246-5261.

[9] MUSS J D, MLADENOFF D J, TOWNSEND P A. A pseudo-waveform technique to assess forest structure using discrete lidar data[J]. Remote Sensing of Environment, 2011, 115(3): 824-835.

[10] 李建更, 罗奥荣, 李晓理. 基于互补集合经验模态分解与支持向量回归的 $PM_{2.5}$ 质量浓度预测[J]. 北京工业大学学报, 2018, 44(12): 1494-1502.
LI J G, LUO A R, LI X L. Prediction of $PM_{2.5}$ mass concentration based on complementary ensemble empirical mode decomposition and support vector regression[J]. Journal of Beijing University of Technology, 2018, 44(12): 1494-1502. (in Chinese)

[11] SUN G, RANSON K J, KIMES D S, et al. Forest vertical structure from GLAS: an evaluation using LVIS and SRTM data[J]. Remote Sensing of Environment, 2008, 112(1): 107-117.

[12] LEFSKY M A, HARDING D J, KELLER M, et al. Estimates of forest canopy height and aboveground biomass using ICESat[J]. Geophysical Research Letters, 2005, 32(22): L22S02.

[13] 胡凯龙, 刘清旺, 庞勇, 等. 基于机载激光雷达校正的 ICESat/GLAS 数据森林冠层高度估测[J]. 农业工程学报, 2017, 33(16): 88-95.
HU K L, LIU Q W, PANG Y, et al. Forest canopy height estimation based on ICESat/GLAS data by airborne lidar[J]. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(16): 88-95. (in Chinese)

[14] 汪垚, 倪文俭, 张志玉, 等. 激光雷达回波模型辅助的坡地森林冠层高度反演[J]. 遥感学报, 2018, 22(3): 98-109.
WANG Y, NI W J, ZHANG Z Y, et al. Retrieval of forest canopy heights by using large-footprint waveform data assisted by the LiDAR model over hillsides[J]. Journal of Remote Sensing, 2018, 22(3): 98-109. (in Chinese)

[15] SCORNET E. Random forests and kernel methods[J]. IEEE Transactions on Information Theory, 2016, 62(3): 1485-1500.

[16] DAVIES A, GHAHRAMANI Z. The random forest kernel and other kernels for big data from random partitions[EB/OL]. [2019-02-20]. https://arxiv.org/pdf/1402.4293.

[17] ARLOT S, GENUER R. Analysis of purely random forests bias[EB/OL]. [2019-02-20]. https://arxiv.org/pdf/1407.3939.

[18] BREIMAN L. Random forests[EB/OL]. [2019-02-25]. https://arxiv.org/pdf/1407.3939.

(责任编辑　梁　洁)