DOI: 10.16452/j.cnki.sdkjzk.2024.06.005 **文章编号**:1672-3767(2024)06-0043-10

基于主动学习的沿海区 GDEM 修正方法

刘盼盼,李艳艳,刘 妍,刘雅婷,陈传法
(山东科技大学 测绘与空间信息学院,山东 青岛 266590)

摘 要:针对传统全球数字高程模型修正方法忽略训练样本质量等问题,提出一种基于主动学习的沿海区全球数字高程模型(GDEM)修正方法。该方法首先选择一定数量的代表性样本点作为初始训练集,然后通过聚类批处理模式采样算法,采用迭代方式选取高质量代表点进行模型训练,最后利用选取的代表点构建机器学习模型以实现 GDEM 修正。以美国克逊维尔为训练区、查尔斯顿为迁移实验区,选取中误差和平均绝对误差验证模型的精度。实验结果表明,与传统的 GDEM 修正方法相比,本研究方法仅需选择 8.57%的采样点即可完成模型训练,且 GDEM 的中误差降低了 3.31%~51.65%、平均绝对误差降低了 4.76%~48.72%。在迁移实验区,修正后 COPDEM30 的中误差从 6.52 m 降至 1.68 m。相比于传统方法,本研究方法的中误差和平均绝对误差分别降低了 24.82%和 30.28%,证明了模型具有一定的迁移性。

关键词:主动学习;全球数字高程模型修正;沿海区
中图分类号:P237 **文献标志码**:A

Global digital elevation model correction method for coastal areas based on active learning

LIU Panpan, LI Yanyan, LIU Yan, LIU Yating, CHEN Chuanfa
(College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: To address issues such as neglecting sample quality in traditional global digital elevation model (GDEM) correction methods, this paper proposed a GDEM correction method for coastal areas based on active learning. Firstly, a certain number of representative sample points were selected as the initial training set. Then, high-quality representative points were iteratively selected for model training through a clustering-based batch processing sampling algorithm. Finally, a machine learning model was constructed by using all selected representative points to achieve GDEM correction. The accuracy of the model was validated by selecting mean square error and mean absolute error with the coastal areas of Jacksonville and Charleston in the United States as the training area and transfer experimental area respectively. Experimental results show that, compared with traditional GDEM correction methods, the proposed method only requires 8.57% of the sampling points to complete model training. The root mean square error of GDEM is reduced by 3.31% to 51.65% and the mean absolute error is reduced by 4.76% to 48.72%. In the transfer experiment area, the root mean square error of the corrected COPDEM30 is reduced from 6.52 m to 1.68 m. Compared with traditional methods, the root mean square error and mean absolute error of the proposed method are reduced by at least 24.82% and 30.28% respectively, demonstrating that the model has a certain level of transferability.

Key words: active learning; global digital elevation model (GDEM) correction; coastal area

收稿日期:2024-02-08
基金项目:国家自然科学基金项目(42271438);山东省自然科学基金项目(ZR2024MD040)
作者简介:刘盼盼(1996—),女,安徽阜阳人,硕士研究生,主要从事空间数据质量改善方面的研究.
E-mail:1428511538 @qq.com
李艳艳(1987—),女,山东潍坊人,副教授,博士,主要从事空间数据质量改善及高频 GNSS 数据处理方面的研究,本文通信作者. E-mail:yylee@sdust.edu.cn

据统计,全球 60%的经济总量、75%的大型城市、70%的工业资本集中在沿海地区[1-2]。然而,受气候变化等因素的影响,海平面上升导致的洪水问题成为沿海地区面临的主要威胁之一。沿海洪水可能引发严重的社会问题,如人员伤亡、财产损失、社会秩序混乱以及环境破坏等,对沿海地区的社会稳定和经济发展造成严重威胁。为了有效预防和减轻洪水灾害的影响,洪水模拟已成为预测和应对洪水事件的关键工具。以遥感技术手段获取的全球数字高程模型(global digital elevation model,GDEM)是洪水模拟的主要输入数据,但因含有大量植被和建筑物等信息,导致现存 GDEM 高程普遍偏高。因此,发展高精度、高时效性的 GDEM 修正技术,对提升洪水模拟分析的准确性和效率,保护沿海地区的安全具有重要意义。

目前,GDEM 修正方法主要分为:直接法、线性回归法和机器学习法。①直接法[3]通过从原始 GDEM 中减去植被高得到修正后的 GDEM。该方法原理简单,但忽略了植被高度的空间异质性带来的预测偏差。②线性回归法[4]通过构建特征因子与 GDEM 误差的线性关系实现 GDEM 修正,但忽略了 GDEM 误差和驱动因素的非线性关系。③相比于前两种方法,机器学习法[5-7]因能较好地拟合自变量和因变量之间复杂的非线性关系而被广泛应用于 GDEM 修正。然而,现有方法在构建普适性机器学习模型时,主要依赖于训练样本的数量,忽略了训练样本的质量,容易导致计算耗时大、模型过拟合等问题。因此,如何准确选择有代表性的样本进而提升机器学习性能是提高 GDEM 修正精度的有效途径。

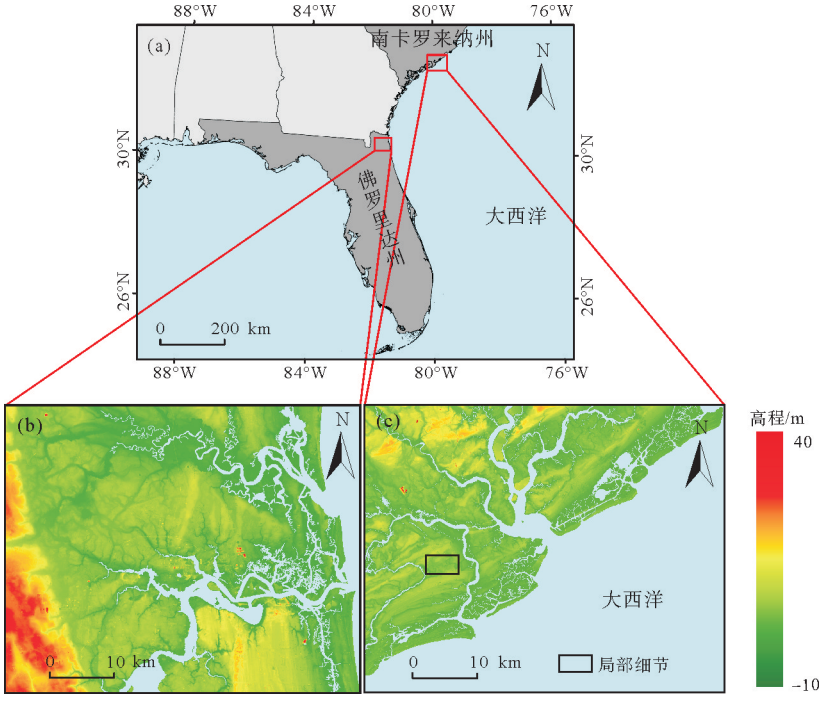
目前,主动学习法因其可自动选择少量高价值训练样本,且能有效提高模型训练速度和模型性能,而被广泛应用于计算机视觉、医学影像分析和自然语言处理等领域[8]。然而,将主动学习应用到 GDEM 修正方面的研究较少。传统的主动学习方法按照每次迭代样本的数量分为逐样本查询法和批处理查询法。①常用的逐样本查询法包括:贪婪采样法(greedy sampling,GS)(如:greedy sampling on the inputs,GSx;greedy sampling on the output,GSy;improved greedy sampling on both inputs and output,iGS)[9]、委员会查询法(query-by-committee,QBC)[10]和模型变化期望最大化法(expected model change maximization,EMCM)[11]。虽然逐样本查询可以更准确地代表数据分布,但耗时巨大,不适用于大规模样本筛选。此外,上述方法只考虑样本的单一特性准则,未能充分顾及样本的多样性、代表性和信息性。②常用的批处理查询法包括:基于池的序贯主动学习回归算法(pool-based sequential active learning for regression,DR)[12]和增强批处理模式主动学习法(enhanced batch-mode active learning,EBMAL)[13],其中 EBMAL 方法虽然充分考虑样本特性、节约了模型训练时间,但在去除离群值时需设置合适的经验阈值,且阈值设定方法过于繁琐和主观。

针对上述瓶颈,本研究提出一种基于主动学习的沿海区 GDEM 修正方法,首先通过对未标记样本聚类,确保所选样本的多样性;再通过异质委员会查询,顾及所选样本的代表性和信息性;最后通过批量选取样本的模式,节约时间,有效选择高质量样本并提升模型精度。

1 研究区概括和实验数据

1.1 研究区概况

本研究选取杰克逊维尔和查尔斯顿沿岸为实验区(图 1),其中前者主要用于模型训练和总体精度验证,后者用于模型迁移性验证。杰克逊维尔(图 1(b))位于佛罗里达州东北部,靠近大西洋海岸,地理位置为 30°16′6″N~30°37′20″N,81°55′5″W~81°23′11″W。地物类型主要包括城市建筑、沿海湿地和自然保护区,地势平坦,高程范围-29~40 m,平均高程 6.6 m,平均坡度 2.9°。查尔斯顿(图 1(c))位于南卡罗来纳州库珀河与温都河交汇处,地理位置为 32°33′13″N~32°56′34″N,80°10′40″W~79°34′26″W,高程范围-16.6~32 m,平均高程 2.87 m,平均坡度 2.4°。此外,本研究选取实验区内激光雷达(light detection and ranging,LiDAR)数字地形模型(digital terrain model,DTM)高程范围-10~80 m 为参考数据。该地区特殊的地理位置更易受到海平面上升、风暴潮和洪水等自然灾害的威胁[14]。



(a)美国部分矢量图;(b)训练区:杰克逊维尔;(c)验证区:查尔斯顿,其中局部细节主要用于后续精度比较

图 1 研究区地理位置和 LiDAR DTM

Fig. 1 Geographical location and LiDAR DTM of the study area

1.2 实验数据

本研究采用的实验数据主要包括机载 LiDAR 点云、30 m 哥白尼数字高程模型(Copernicus DEM GLO-30,COPDEM30)、植被数据、建筑物数据、夜间灯光和 Landsat 8 等(表 1),原因是地表覆盖类型、冠层高度和植被覆盖率等植被分布和结构因素对地表的反射率和高程信息有显著影响,且建筑物高度、面积、体积以及人口密度数据是反映地表人类活动的关键因素,夜间灯光数据能够反映城市和人类活动的分布情况,可提供有关建筑物使用和人口密集度的信息。Landsat 8 提供了高质量的地表影像数据,其多波段信息能够捕捉地表特征的变化,可全面获取地表特征信息。

表 1 数据来源与数据类型

Table 1 Data source and data type

基础数据	数据来源	数据类型	数据年份	分辨率/m
机载 LiDAR[15]	3DEP LidarExplorer https://apps.nationalmap.gov/lidar-explorer/#/	矢量	2014	—
COPDEM30[16]	Opentopography 网站 https://portal.opentopography.org/dataCatalog?group=global	栅格	2010—2015	30
植被数据[17-18]	全国地理信息资源目录服务系统 https://www.webmap.cn/mapDataAction.do?method=globalLandCover	栅格	2010	30
	美国航天局喷气推进实验室 http://lidarradar.jpl.nasa.gov	栅格	—	1 000
	GLAD https://glad.umd.edu/dataset/	栅格	2010	30
建筑物和人口密度数据[19]	欧洲联盟委员会 https://ghsl.jrc.ec.europa.eu/	栅格	2010—2018	100
夜间灯光[20]	美国国家地球物理数据中心 https://eogdata.mines.edu/products/vnl/#annual_v2	栅格	2015	500
Landsat 8[21]	美国地质调查局 https://earthexplorer.usgs.gov/	栅格	—	30/15

1.2.1 COPDEM30

哥白尼数字高程模型(copernicus digital elevation model,COPDEM)由欧洲航天局(European space agency,ESA)发布,从 WorldDEM 数据集中获得,并借助其他高程数据,如高级陆地观测卫星全球数字表面模型(ALOS World 3D-30 m,AW3D30)、先进星载热发射和反射辐射仪全球数字高程模型(advanced spaceborne thermal emission and reflection radiometer global digital elevation model,ASTER DEM)以及航天飞机雷达地形测绘使命数字高程模型(shuttle radar topography mission digital elevation model,SRTM DEM)等局部填充得到。COPDEM 有 3 种不同分辨率产品:10、30 和 90 m。其中,10 m 仅覆盖 39 个欧洲国家,30 和 90 m 覆盖全球。COPDEM30 在 90%线概率误差(line error at 90% probability,LE90)的总体绝对高程精度优于 4 m,在 90%圆概率误差(circular error at 90% probability,CE90)的绝对平面精度优于 6 m。根据 COPDEM 产品覆盖范围和分辨率,本研究选取 30 m COPDEM 作为研究对象。

1.2.2 机载 LiDAR DTM

机载 LiDAR 点云以高精度、高密度、高植被穿透力等优点,广泛应用于高精度 DTM[22]。本研究从公共数据源 3DEP LidarExplorer 网站获取实验区 LiDAR 点云数据,研究区域 LiDAR 点云数据参数描述如表 2 所示。为了获得测区地面点云,首先使用郭娇娇等[23]提出的滤波方法获取初始地面点,其次采用人工编辑方式对滤波点云进一步处理以获得高精度地面点云,然后使用克里金插值法对该地面点插值生成 30 m 分辨率的 DTM,最后提取高程为−10～80 m 的栅格(共计 1 667 320 个网格)。

表 2 机载 LiDAR 点云参数描述

Table 2 Parameters for the collection of airborne LiDAR point cloud data

区域	扫描设备	飞行高度/m	点密度/(pts/m^2)	地面点密度/(pts/m^2)	脉冲频率/kHz	垂直精度/cm
杰克逊维尔	Leica ALS80	1 400	24.64	1.88	586	5.7
查尔斯顿	Riegl LMS-Q1560	2 043	7.72	0.81	800	＜6.0

2 基于聚类批处理模式主动学习的沿海区 GDEM 修正方法

针对基于传统机器学习方法的 GDEM 修正方法容易忽略训练样本质量等问题,提出一种基于聚类批处理模式主动学习的沿海区 GDEM 修正方法,主要步骤包括数据预处理、模型构建、GDEM 修正和精度评价(图 2)。

2.1 数据预处理

由于 GDEM 与 LiDAR DTM 的坐标系统不统一,首先,借助美国国家海洋和大气管理局(national oceanic and atmospheric administration,NOAA)开发的 VDatum(https://vdatum.noaa.gov/),将其转换到 WGS84/EGM2008;然后,采用 Nuth 等[2]提出的配准方法,以 LiDAR 为参考对 COPDEM30 进行配准;再在 ArcGIS10.5 中提取高程范围在−10～80 m 的沿海区 GDEM,并从 GDEM 中提取坡度、坡向和地形起伏度等地形因子;最后,根据 LiDAR 数据提取对应点的特征因子并将样本点随机分成 7∶3,其中 70%(共计 1 167 124 个)作为初始训练点,用于主动学习模型样本筛选,剩余 30%(共计 500 196 个)用于模型精度验证和 GDEM 精度评价。

2.2 模型构建

针对传统主动学习方法容易导致样本冗余、模型性能不佳等问题,提出一种基于聚类批处理模式主动学习(cluster-based batch mode active learning,CBMAL)的 GDEM 修正方法。该方法主要包括 3 部分:初始样本点选择、样本迭代选择和 GDEM 修正,具体过程如图 2 所示。

2.2.1 初始训练点选择

初始样本点一般通过随机或多样性抽样策略实现。随机抽样尽管操作简便,但可能造成训练集代表性不足。相比之下,多样性抽样以样本间相似度或距离为依据选取一组代表性样本,以提高初始样本的

多样性,从而提升模型性能。然而,该方法计算复杂,且易受到特征空间维度的影响,不适用于处理大规模数据集。

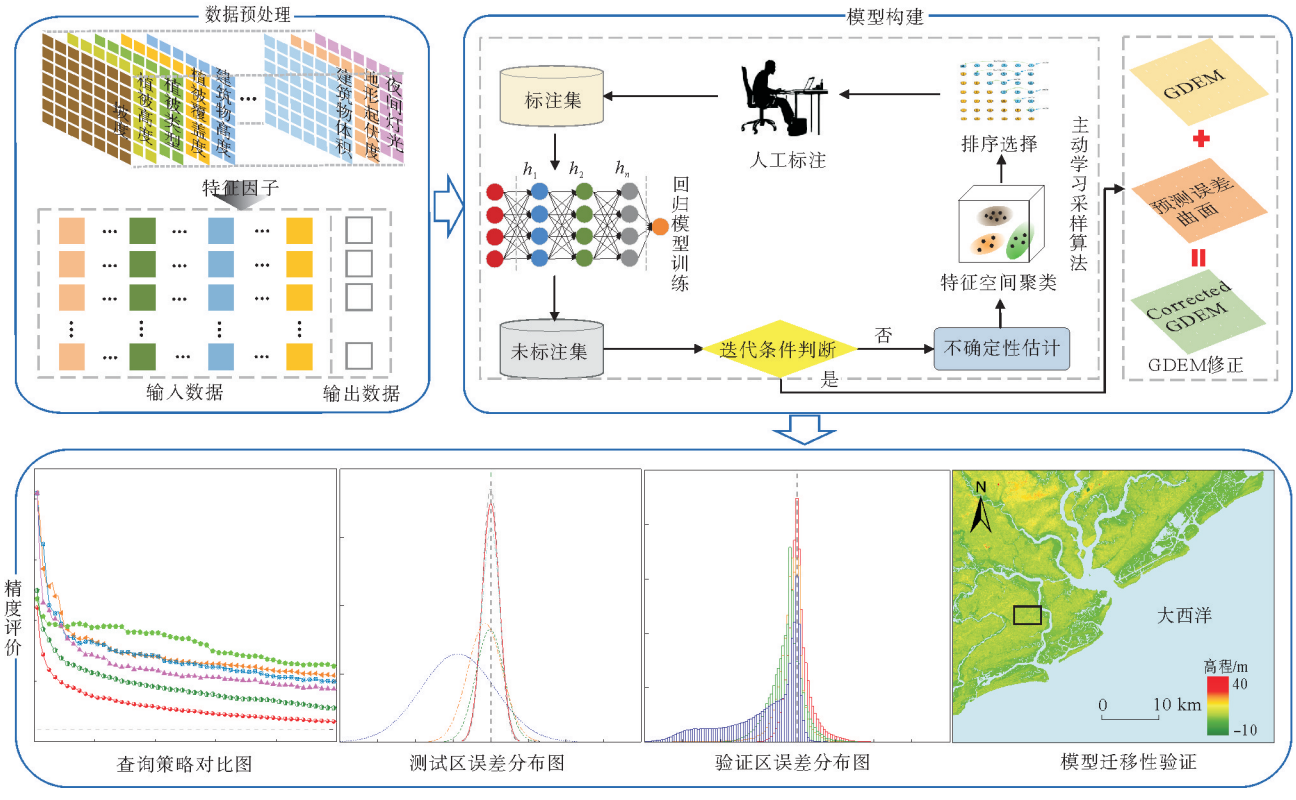


图 2　基于聚类批处理模式主动学习的沿海区 GDEM 修正方法过程图

Fig. 2　Process diagram of GDEM correction method in coastal areas based on active learning of clustering batch mode

为此,为充分顾及所选初始样本的代表性和多样性,通过聚类簇质心的方法选取初始样本点。具体步骤为:①定义模型的未标注样本集 $D=\{(x_i,y_i)i=1,2,\cdots,n\}$,其中 x_i 为第 i 个样本的特征变量,y_i 为第 i 个样本的输出,n 为未标注集样本数量;②对研究区未标注样本集执行 k 均值聚类($k=m$,m 为每次迭代选取未标注点的数量);③选取每个聚类簇中距离质心最近的样本作为初始训练点添加到标注样本集,并将剩余样本点作为未标注点集用于主动学习模型的迭代。该过程确保了初始标注样本的代表性和多样性,为建立高质量机器学习模型奠定了基础。

2.2.2　样本迭代选择

传统委员会查询方法通常利用自举法(bootstrap)构建同质委员会,即通过对标记数据进行有放回的抽样形成多个独立的训练样本,进而训练多个相同类型或结构的回归模型。这种方法确保了每个模型在训练过程中的差异性,从而提高了整体预测性能的鲁棒性。然而,由于所有模型的结构和算法相同,导致该方法在结构和算法上缺乏本质性差异,容易在模型训练过程中产生相似的偏差,未能充分顾及样本的多样性、代表性和信息性。为此,提出一种结合聚类和委员会查询思想的方法,该方法采用异质委员会查询代替传统的同质委员会查询,旨在从不同层面捕捉数据特征,从而提升主动学习模型的效率和训练样本的质量。具体步骤如下:

1) 基于初始样本点构建一个学习委员会,该委员会由多个异质性基学习器组成,包括随机森林(random forest,RF)、多层感知器(multilayer perceptron,MLP)和支持向量回归(support vector regression,SVR)。

2) 采用已有标记样本集分别训练委员会模型,并利用训练好的模型对每个未标注样本点进行预测。

3) 利用式(1)计算每个未标记样本点的预测方差,该方差反映了各委员会模型对未标注样本预测结果的差异程度,从而量化未标注样本的不确定性。

4) 对未标注集进行 k 均值聚类,将其划分为 m 个聚类簇,每个簇代表数据集中一个独特的区域或特征以确保所选样本的多样性。

5) 在每个聚类簇中,根据样本的预测方差进行排序,其中方差越大表明该样本的预测越不确定、信息性越高,因此选择每个聚类簇中方差最大的样本进行人工标注,并更新标注集和未标注集。

6) 重复步骤 2)~5)直至达到预设的标注点总数或训练精度,通过上述步骤,能够综合考虑样本的多样性、代表性和信息性,有效提高主动学习的效率和模型性能。

$$\sigma_n=\frac{1}{N}\sum_{j=1}^{1}(y_i^j-\bar{y}_i)^2,(i=n-m,n-m+1,\cdots,n)。\tag{1}$$

式中,$\bar{y}_i=\frac{1}{N}\sum_{j=1}^{N}y_i^j$,$N$ 代表模型委员会个数,$y_i{}^j$ 表示第 j 个回归模型对未标注样本 $\{x_i\}_{i=n-m}^{n}$ 的预测值。

2.3 GDEM 修正

根据上述过程选取的训练样本,借助 RF 模型捕捉驱动因子与 GDEM 误差之间的关系,并利用训练好的模型对 GDEM 修正。修正模型表达式为:

$$E=f(S_{\mathrm{L}},T_{\mathrm{F}},V_H,V_C,B_{\mathrm{H}},B_{\mathrm{A}},B_{\mathrm{V}},P_{\mathrm{op}},V_{\mathrm{I}},H_0,H_1,H_2,\cdots,H_8)。\tag{2}$$

式中:$E=H_{\mathrm{L}}-H_{\mathrm{G}}$,$H_L$ 表示 LiDAR 点高程,H_{G} 表示对应的 GDEM 高程;E 表示 GDEM 误差;S_{L} 为坡度;T_{F} 为地形起伏度;H_0 为待修正点高程;$H_1,H_2,\cdots,H_8$ 为空间邻域;V_{H} 为植被高度;V_{C} 为植被覆盖度;V_{T} 为植被类型;B_{H} 为建筑物高度;B_{A} 为建筑物面积;B_{V} 为建筑物体积;P_{op} 为人口数据;V_{I} 为夜间灯光数据。

2.4 对比方法与性能评价指标

为验证本研究方法的精度,首先将训练区(杰克逊维尔)样本点随机分成 7∶3,其中前者用于模型训练,后者用于模型精度验证;然后将训练好的模型应用于验证区(查尔斯顿),以检验模型的迁移性。

为分析本研究方法的高效性,将本研究算法与传统的主动学习算法 GS(GSx、GSy、iGS)[9]、QBC[10]和 DR[12]对比分析。其中,GSx、GSy 和 iGS 分别是在输入、输出和坐标空间中通过选择与标记点最远的一批未标记点进行标记;QBC 是根据现有的标记数据集建立一个学习者委员会(通常采用自举法),然后迭代选取委员会不确定性最大的未标记样本进行标记;DR 采用聚类算法对样本进行聚类,并在每次迭代过程中逐步增加聚类簇的数量。每次迭代中,先排除包含已标注样本的聚类簇,再对剩余聚类簇按其包含的样本数量进行排序。最后从样本数量最多的聚类簇中,选择距离该簇质心最近的样本进行标注。此外,为避免迭代次数过多和计算机内存不足问题,设置每次迭代选择的采样数为 1 000。将本研究计算结果与三种近期公开修正后的 GDEM 进行精度比较。这三种 GDEM 分别是 Kulp 等[5]生成的修正后沿海区 GDEM(coastal digital elevation model,CostalDEM v1.1)、Hawker 等[6]生成的去除建筑物和植被的 GDEM(forest and buildings removed copernicus DEM,FABDEM)和 Dusseau 等[7]生成的高程为-10~80 m 的全球沿海区数字高程模型(DuilvumDEM)。

选用平均绝对误差 M_{AE} 和均方根误差 R_{MSE} 作为精度评价因子,计算公式分别为:

$$M_{\mathrm{AE}}=\frac{\sum_{i=1}^{n}|L_i-G_i|}{n},\tag{3}$$

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(L_i-G_i)^2}。\tag{4}$$

式中:L_i 和 G_i 分别表示第 i 个 LiDAR 点高程和对应的 GDEM 点高程,n 为检核点个数。

3 实验结果及分析

3.1 主动学习方法对比

图 3 展示了 7 种主动学习方法的中误差与训练点数的关系。结果表明,随着训练点数的增加,7 种算法的计算精度均明显提高。但当训练点数增加到一定程度后,各方法计算精度的提升幅度均逐渐减缓。原因是随着训练样本数量的增加,机器学习模型的性能趋于饱和。7 种算法在相同采样点数下,本研究方法(CBMAL)精度最优,DR 次之,QBC 最差,原因是本研究方法不仅在聚类算法中考虑了样本点的多样性和代表性,还通过方差计算增强了样本的信息性。但由于本研究设置的每次迭代采样数为 1 000,降低了 GSx、GSy 和 iGS 样本间的多样性,造成样本冗余;而 QBC 虽然顾及样本的信息性和代表性,但忽略了样本间的多样性。相比之下,DR 在聚类算法中同时考虑了样本的代表性和多样性,但忽略了样本的信息性。需要注意的是,当训练点数量达到 100 000 时(即训练点约为原始点数的 8.57%),本研究方法的计算精度与基于全部样本点,即未经筛选的训练样本点(简称全部点)的计算精度非常接近,且与 GSx、GSy、iGS、QBC 和 DR 方法相比,本研究方法的 R_{MSE} 分别降低了 24.67%、22.15%、18.89%、28.39%和 8.66%。

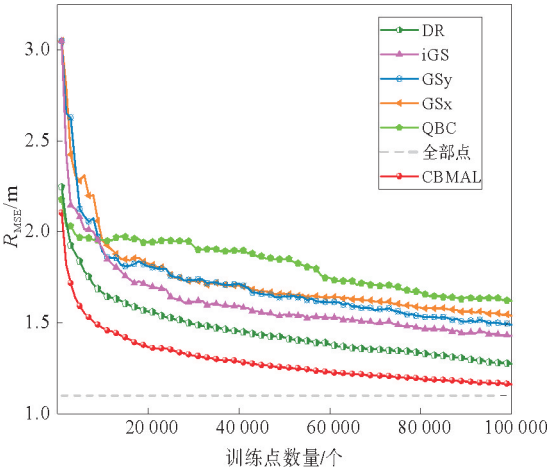


图 3 不同训练点数对各主动学习方法计算精度的影响

Fig. 3 Effect of the number of training points on the RMSEs of the active learning methods

3.2 整体精度分析

图 4 为 GDEM 修正前后的误差统计,可以看出修正后 COPDEM30 精度(R_{MSE}=1.17 m,M_{AE}=0.8 m)优于修正前(R_{MSE}=6.58 m,M_{AE}=4.5 m),原因是原始 COPDEM30 中含有大量植被和建筑物高度信息,导致真实地表高程被普遍高估。此外,相比 CostalDEM v1.1(R_{MSE}=1.87 m,M_{AE}=1.37 m)、FABDEM(R_{MSE}=2.42 m,M_{AE}=1.56 m)和 DiluviumDEM(R_{MSE}=1.21 m,M_{AE}=0.84 m),本研究方法精度最优。具体而言,本研究方法的 M_{AE} 相较于 DiluviumDEM、CostalDEM v1.1 和 FABDEM 分别降低了 4.76%、41.61%、48.72%,R_{MSE} 分别降低了 3.31%、37.43%和 51.65%。但本研究方法的计算精度略低于基于全部训练样本点 GDEM 的精度,原因是前者仅采用后者约 8.57%的训练点。

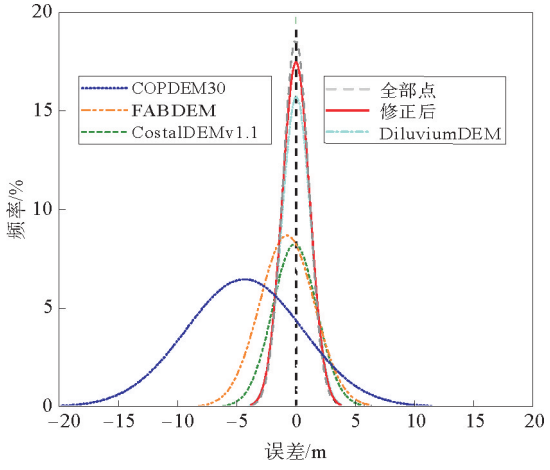


图 4 测试区 GDEM 修正前后误差统计

Fig. 4 Error statistics of GDEM before and after correction in the experimental area

3.3 模型迁移性验证

为了验证模型的迁移性,本研究将杰克逊威尔(训练区)训练好的模型应用于查尔斯顿(验证区)。结果(图 5)表明,修正前 COPDEM30 的误差集中分布在−20~3 m 之间,说明 COPDEM30 存在负偏差。本研究方法修正后 GDEM(R_{MSE}=1.68 m,M_{AE}=1.12 m)精度明显优于 COPDEM30(R_{MSE}=6.52 m,M_{AE}=4.44 m)、CostalDEM v1.1(R_{MSE}=2.68 m,M_{AE}=1.96 m)和 FABDEM(R_{MSE}=2.31 m,M_{AE}=1.43 m)。

具体而言,与 CostalDEM v1.1 和 FABDEM 相比,本研究方法的 M_{AE} 分别降低了 42.86%和 21.68%,R_{MSE} 分别降低了 37.31%和 27.27%。

图 6 展示了 COPDEM30 修正前后以及与 LiDAR DTM 的地形细节对比。由图 6 可以看出,ICOPDEM30(图 6(b))相较于 COPDEM30(图 6(a)),在高程分布上更接近于 LiDAR DTM(图 6(c)),且更有效地保持了地形特征。由 GDEM 误差图可知,原始 COPDEM30 相较于 LiDAR DTM 存在显著的正偏差,原因是 COPDEM30 中包含了大量建筑物和植被高度信息,而本研究方法修正后 GDEM 误差更趋近于零,特别是在图 6 中的椭圆区域内。此外,GDEM 的局部细节图也表明,本研究方法修正后的 GDEM 更贴近 LiDAR DTM,能够真实反映地表形态。因此,本研究方法不仅提高了 COPDEM30 精度,而且更好地保持了地形特征信息。

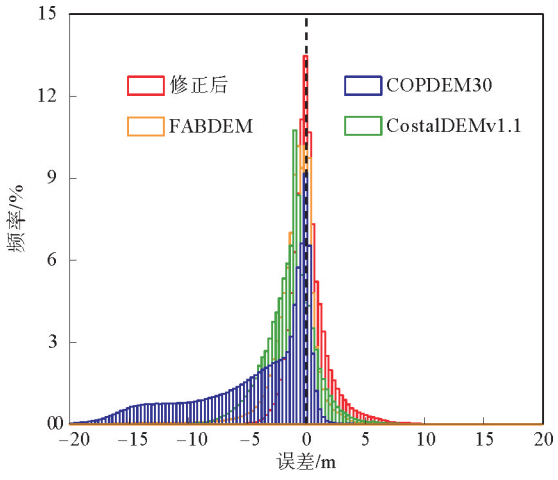
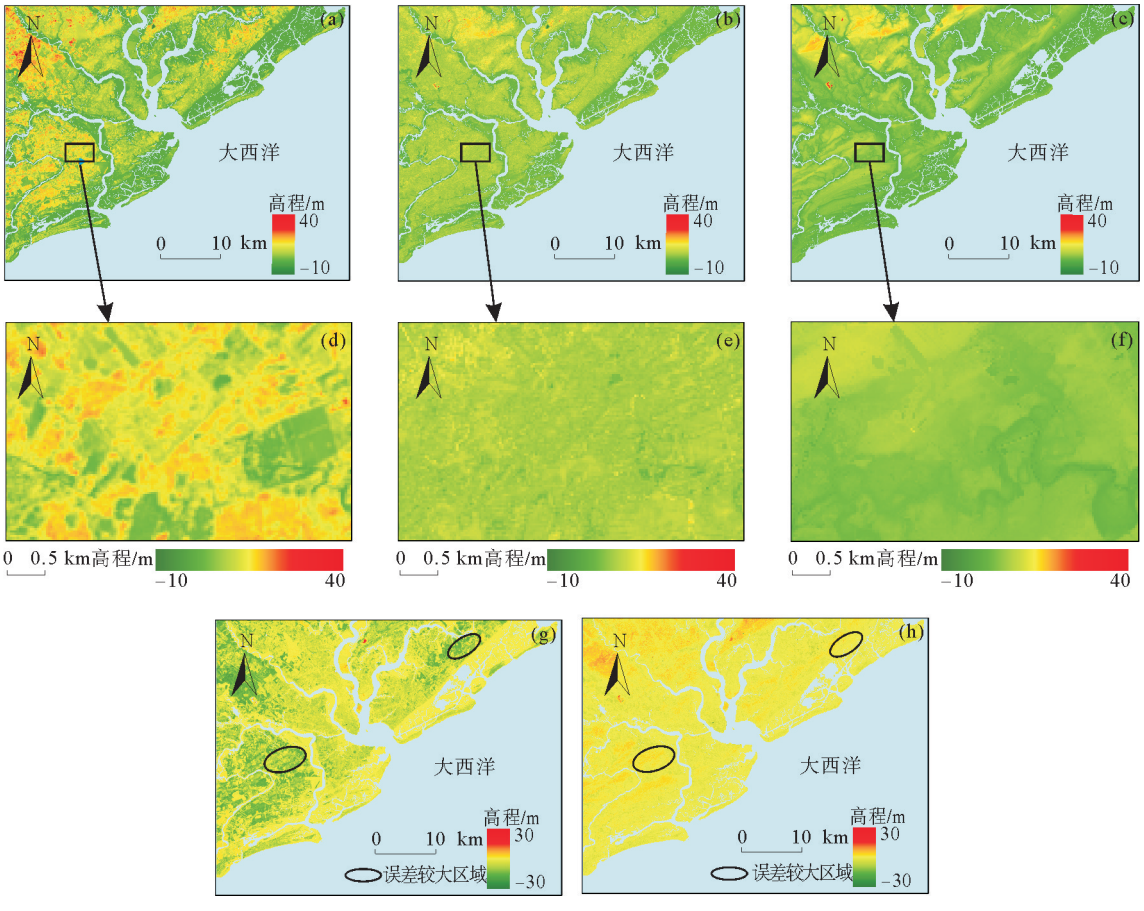


图 5 验证区 COPDEM30 修正前后误差直方图(LiDAR DTM-GDEM)

Fig. 5 Error histogram of COPDEM30 before and after correction(LiDAR DTM-GDEM)



(a) COPDEM30;(b) ICOPDEM30;(c) LiDAR DTM;(d) COPDEM30 局部细节图;(e) ICOPDEM30 局部细节图;(f) LiDAR DTM 局部细节图;(g) LiDAR DTM-COPDEM30;(h) LiDAR DTM-ICOPDEM30

图 6 COPDEM30 修正前后与 LiDAR DTM 对比

Fig. 6 Comparison of COPDEM30 before and after correction and LiDAR DTM

4 结论

为提升沿海 GDEM 产品精度,提出一种基于主动学习的沿海区 GDEM 修正方法。该方法首先采用聚类批处理模式主动学习法挑选一定数量的高质量代表性样本点,然后基于这些样本点构建 GDEM 修正模型,最后将训练好的模型用于 COPDEM30 修正。实验结果表明:

1) 相较于传统主动学习方法,本研究方法的 R_{MSE} 至少降低了 8.66%,表明本研究方法能够准确挑选出高质量代表性样本点。

2) 修正后的 GDEM(DiluviumDEM、CostalDEMv1.1 和 FABDEM)相较于传统方法精度最高,说明样本质量对模型预测能力的影响不容忽视。

3) 在迁移实验区,本研究方法修正后 GDEM 精度优于其他的方法,表明本研究方法具有一定的泛化能力。

参考文献:

[1] MUIS S,VERLAAN M,WINSEMIUS H C,et al. A global reanalysis of storm surges and extreme sea levels[J/OL]. Nature Communications,2016,7. DOI:10.1038/ncomms11969.

[2] NUTH C,KÄÄB A. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change[J]. The Cryosphere,2011,5(1):271-290.

[3] O'LOUGHLIN F E,PAIVA R C D,DURAND M,et al. A multi-sensor approach towards a global vegetation corrected SRTM DEM product[J]. Remote Sensing of Environment,2016,182:49-59.

[4] SU Y J,GUO Q H. A practical method for SRTM DEM correction over vegetated mountain areas[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2014,87:216-228.

[5] KULP S A,STRAUSS B H. CoastalDEM:A global coastal digital elevation model improved from SRTM using a neural network[J]. Remote Sensing of Environment,2018,206:231-239.

[6] HAWKER L,UHE P,PAULO L,et al. A 30 m global map of elevation with forests and buildings removed[J/OL]. Environmental Research Letters,2022,17. DOI:10.1088/1748-9326/ac4d4f.

[7] DUSSEAU D,ZOBEL Z,SCHWALM C R. DiluviumDEM:Enhanced accuracy in global coastal digital elevation models[J/OL]. Remote Sensing of Environment,2023,298. DOI:10.1016/j.rse.2023.113812.

[8] 邹禄杰,花向红,赵不钒,等. 点云场景语义标注的排序批处理模式主动学习法[J]. 测绘学报,2022,52(2):260-271.
ZOU Lujie,HUA Xianghong,ZHAO Buftan,et al. Ranked batch-mode active learning method for semantic annotation of point cloud scene[J]. Acta Geodaetica et Cartographica Sinica,2022,52(2):260-271.

[9] WU D,LIN C T,HUANG J. Active learning for regression using greedy sampling[J]. Information Sciences,2019,474:90-105.

[10] BURBIDGE R,ROWLAND J J,KING R D. Active learning for regression based on query by committee[C]//Intelligent Data Engineering and Automated Learning-IDEAL 2007. Birmingham:Springer. 2007:209-218.

[11] CAI W B,ZHANG Y,ZHOU J. Maximizing expected model change for active learning in regression[C]//2013 IEEE 13th International Conference on Data Mining. Shanghai,China:IEEE. 2013:51-60.

[12] WU D R. Pool-based sequential active learning for regression[J/OL]. IEEE transactions on Neural Networks and Learning Systems,2018,30. DOI:10.1109/TNNLS.2018.2868649.

[13] WU D R,LAWHERN V J,GORDON S,et al. Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning(EBMAL) for regression[C]//2016 IEEE International Conference on Systems,Man,and Cybernetics (SMC). Budapest:IEEE,2016:730-736.

[14] BARBIER E B. A global strategy for protecting vulnerable coastal populations[J]. Science,2014,345(6202):1250-1251.

[15] ARUNDEL S T,PHILLIPS L A,LOWE A J,et al. Preparing the national map for the 3D elevation program:Products,process and research[J]. Cartography and Geographic Information Science,2015,42(S1):S40-S53.

[16] FRANKS S,RENGARAJAN R. Evaluation of Copernicus DEM and comparison to the DEM used for Landsat Collection-2 processing[J/OL]. Remote Sensing,2023,15. DOI:10.3390/rs15102509.

[17] HESS L L,MELACK J M,NOVO E M L M,et al. Dual-season mapping of wetland inundation and vegetation for the central Amazon basin[J]. Remote Sensing of Environment,2003,87:404-428.

[18] SIMARD M,PINTO N,FISHER J B,et al. Mapping forest canopy height globally with spaceborne lidar[J/OL]. Journal of Geophysical Research,2011,116. DOI:10.1029/2011jg001708.

[19] FLORCZYK A J,CORBANE C,EHRLICH D,et al. GHSL data package 2019[M]. Luxembourg:Publications Office of the European Union,2019.

[20] ELVIDGE C D,ZHIZHIN M,GHOSH T,et al. Annual time series of global VIIRS nighttime lights derived from monthly averages:2012 to 2019[J/OL]. Remote Sensing,2021,13. DOI:10.3390/rs13050922.

[21] LOVELAND T R,IRONS J R. Landsat 8:The plans,the reality,and the legacy[J/OL]. Remote Sensing of Environment,2016,185. DOI:10.1016/j.rse.2016.07.033.

[22] 宿殿鹏,阳凡林,陈亮,等 无人机载 LiDAR 测深系统进行海岸带测绘的可行性分析[J]. 山东科技大学学报(自然科学版),2022,41(5):11-20.
SU Dianpeng,YANG Fanlin,CHEN Liang,et al. Feasibility analysis of UAV-airborn LiDAR bathymetry system for costal zone mapping[J]. Journal of Shandong University of Science and Technology(Natural Science),2022,41(5):11-20.

[23] 郭娇娇,陈传法,姚喜,等. 基于多特征聚类的复杂环境机载点云层次滤波方法[J]. 测绘学报,2023,52(10):1724-1737.
GUO Jiaojiao,CHEN Chuanfa,YAO Xi,et al. A multi-feature clustering-based hierarchical filtering method for airborne LiDAR point clouds in complex landscapes[J]. Acta Geodaetica et Cartographica Sinica,2023,52(10):1724-1737.

(责任编辑:高丽华)