

一种基于地形因素的空气质量空间估计方法

吕明琪 李一帆 陈铁明

(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 空气质量监测对于污染评估、危害降低和环保治理等具有重要的指导意义。然而,由于空气质量监测站点的数量十分有限,且空气质量随位置的变化是非线性的,因此空气质量空间估计(即估计任意无空气质量监测站点位置的空气质量)是一项具有挑战性的工作。目前最先进的空气质量空间估计方法考虑了交通、人流、POI 等因素,并基于机器学习技术建立估计模型。然而,这些方法仍存在如下不足:1)由于考虑的因素主要反映城区的特性,因此只能局限在城区范围内使用;2)直接使用从各类因素中提取的特征建立模型,没有对特征进行更深层次的提炼。针对上述问题,提出了一种基于地形因素的空气质量空间估计方法。在该方法中,首先建立地形数据库并提取地形特征,然后基于集成决策树模型对地形特征进行深层转换,最后基于因子分解机建立回归模型。基于真实数据的实验表明,该方法对估计自然地形(如高原、森林、水域等)区域中的空气质量有明显的优势。

关键词 空气质量指数,空间估计,地形因素,集成决策树模型

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.041

Spatial Estimation Method of Air Quality Based on Terrain Factors

LV Ming-qi LI Yi-fan CHEN Tie-ming

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Monitoring air quality is important for pollution evaluation, harm reduction and environmental protection. However, since the number of air quality monitoring stations is extremely limited and air quality varies non-linearly with the change of location, spatial estimation of air quality (i. e., estimating the air quality of any location without an air quality monitor station) is a challenging task. Currently, the most state-of-the-art spatial estimation methods for air quality take the factors such as traffic flow, human mobility and POI into account, and build estimation models based on machine learning. However, there are still some limitations in these methods. On one hand, the considered factors mainly reflect the characteristics of urban area, so these methods are constrained to be used in urban area. On the other hand, these methods train models based on features directly extracted from the factors without refinement. Aiming at these problems, this paper proposed a spatial estimation method of air quality based on terrain factors. First, terrain database is established and terrain features are extracted. Then, the original terrain features are deeply converted based on an ensemble decision tree model. Finally, a regression model is trained based on factorization machine. The experiments on real datasets suggest that the proposed method has advantage in terms of estimating the air quality over the areas with natural terrain (e. g., highland, forest, water).

Keywords Air quality index, Spatial estimation, Terrain factor, Ensemble decision tree model

1 引言

空气污染是一个严重的社会问题,监测空气质量对于污染评估、危害降低和环保治理等具有重要的指导意义^[1]。目前,对空气污染进行量化评价的常用指标为空气质量指数(AQI),而AQI是通过空气质量监测站点进行监测的,空气质量监测站点的建设和维护成本很高(包括建设资金、占地面积、维护人力等),导致空气质量监测站点的数量十分有限(例如,截至2016年,全国城市平均空气质量监测站点的数量约

为6个),这使得AQI数据在空间上存在极大的稀疏性。因此,空气质量空间估计是指估计任意位置的AQI(无论该位置是否有空气质量监测站点)。现有的空气质量空间估计方法主要包括基于空间插值的方法和基于机器学习的方法。

基于空间插值的方法(如最近邻插值法、反距离插值法^[2]、克里金插值法^[3]等)主要考虑当前位置与周边空气质量监测站点间的空间距离,对周边空气质量监测站点的AQI真实值进行线性加权求得AQI估计值。然而,由于空气质量受多种因素(如天气、污染源等)影响,其随空间位置的变化是非

到稿日期:2017-11-04 返修日期:2018-02-13 本文受浙江省自然科学基金(LY18F020033),国家自然科学基金浙江两化融合联系基金(U1509214),国家自然科学基金(61772026)资助。
吕明琪(1982—),男,博士,副教授,CCF会员,主要研究方向为普适计算,E-mail:mingqilv@zjut.edu.cn;李一帆(1992—),男,硕士生,主要研究方向为城市计算;陈铁明(1978—),男,博士,教授,CCF会员,主要研究方向为网络安全,E-mail:tmchen@zjut.edu.cn(通信作者)。

线性的^[4],导致基于空间插值的方法的准确度较低。

基于机器学习的方法综合考虑对空气质量产生影响的各类因素,训练得到可对 AQI 进行非线性空间估计的模型。例如,Zheng 等^[4]综合考虑各类静态因素(如路网结构、POI 分布)和动态因素(如天气、交通流),分别使用条件随机场和人工神经网络建立子模型,并对子模型进行协同训练得到空气质量空间估计模型 U-Air。Chen 等^[5]在 U-Air 的基础上考虑人口流因素,对多个子模型进行集成得到空气质量空间估计模型。然而,由于子模型集成方法仅适用于分类问题,上述方法估计的是 AQI 等级,而非 AQI 数值,因此估计结果的粒度过粗。

为估计 AQI 数值,许多现有方法基于回归算法建立空气质量空间估计模型,其中使用得最广泛的模型为土地利用回归模型(Land Use Regression)^[6-7],该方法首先引入交通、人口、土地利用等因素,然后采用回归模型分析这些因素与空气质量间的关系。此外,Jutzeler 等^[8]对城市进行分区,然后采用高斯过程回归(Gaussian Process Regression)估计区域内的空气质量。Hasenfratz 等^[9]使用非线性的广义加性模型(Generalized Additive Model)对空气质量进行空间估计。

然而,现有基于回归模型的 AQI 空间估计方法仍存在以下不足:1)没有考虑地形因素(如森林、湖泊、高原等)对空气质量的影响,因此通常局限在城市范围内使用,在自然地形区域难以取得较好的效果;2)直接使用从各类基础数据中提取的原始特征建立模型,没有对原始特征进行更深层次的提炼。

针对上述问题,本文提出了一种基于地形特征的 AQI 空间估计方法。该方法利用有限空气质量监测站点的 AQI 真实值以及天气和地形数据建立回归模型,以估计任意无空气质量监测站点位置的 AQI 数值。与现有 AQI 空间估计方法不同,该方法的创新在于:1)利用地形数据提取地形特征(如当前位置地形分布、当前位置与周边空气质量监测站点的地形相似度等),用于建立回归模型;2)基于集成决策树模型对原始特征进行提炼,形成更深层次的特征,用于提升回归模型的性能。

2 方法

2.1 地形数据库

地形地貌是影响空气质量的重要因素^[10],两个周边地形分布差异巨大的位置,即使在空间上很接近,其空气质量的变化规律也通常存在很大差异。例如,图 1 给出了两个空气质量监测站点的 AQI 变化规律。

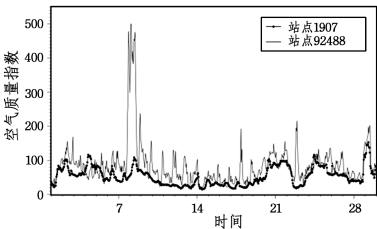


图 1 地形因素对空气质量的影响

Fig. 1 Influence of terrain factors on air quality

图 1 中,站点 92488 部署在黄山市区范围内,站点 1907 部署在千岛湖附近,其空间距离约为 67.7 km。可以看出,这

两个站点虽然在空间上很接近,但其 AQI 变化曲线却有着很大的差异,站点 1907 处的空气质量明显优于站点 92488 处的空气质量。

为建立地形数据库,首先对 OpenStreetMap 地图进行解析。OpenStreetMap 是一个基于网络众包模式的开源、可编辑的地图,包含了大量由用户贡献的地理形状和标签。通过对 OpenStreetMap 地图的解析,可得到大量地理形状和反映其类型的标签。利用标签,保留其中属于自然地形的地理形状,并将这些地理形状粗略地分成森林、水域、沙漠三大类(见图 2)。基于 Google Elevation API 可获得当前位置的海拔。某一区域的地形数据包括该区域的自然地形分布情况和海拔。

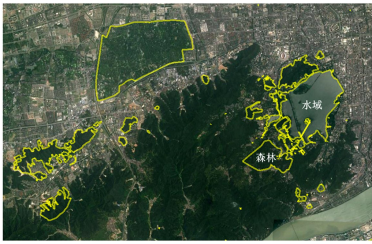


图 2 自然地形地理形状示例

Fig. 2 Example of natural terrain geographic shapes

2.2 原始特征

由于空气质量受多种因素影响,而现有工作从大量城市基础数据中提取特征,如路网数据、出租车数据、交通数据、社交网络数据等^[4-5],这些数据的覆盖率十分有限,导致现有方法的适用范围更加局限。与现有工作不同,本文从具有高覆盖率的数据中提取特征,包括空气质量数据、地形数据、气象数据和时间数据。

2.2.1 气象和时间特征

根据现有研究,气象因素对空气质量有较大影响^[4]。因此,给定位置 l ,提取 l 所在地区的天气 $l.w$ (包括晴、阴、雨、雪、多云和其他),温度 $l.t$ 和风力等级 $l.f$ 作为气象特征。此外,提取季节 s (包括春、夏、秋、冬),是否为工作日 w (包括工作日、休息日),时间段 t (包括白天、晚上)作为时间特征。

最终的气象和时间特征向量为 $mv(l)=\langle l.w, l.t, l.f, s, w, t \rangle$ 。其中,所有类别型特征(包括 $l.w, s, w, t$)采用 one-hot 编码进行表示,所有数值型特征(包括 $l.t, l.f$)采用 min-max 归一化进行处理,则气象和时间特征向量的维度为 16。

2.2.2 本地地形特征

本地地形特征用于表示待估计位置周边的地形。给定位置 l ,首先计算 l 的本地地形分布特征,计算步骤如下:1)以 l 为中心构造一个边长为 d 的正方形区域 $l.G$;2)查询获取 $l.G$ 与其周围自然地形地理形状的交叠部分,并分别计算与森林地形地理形状交叠部分的面积 $l.A_f$ 、与水域地形地理形状交叠部分的面积 $l.A_w$ 和与沙漠地形地理形状交叠部分的面积 $l.A_s$;3)计算森林地形面积占比 $l.p_f=l.A_f/d^2$ 、水域地形面积占比 $l.p_w=l.A_w/d^2$ 和沙漠地形面积占比 $l.p_s=l.A_s/d^2$,则普通地形面积占比为 $l.p_n=1-l.p_f-l.p_w-l.p_s$,然后基于 Google Elevation API 获得 l 处的海拔 $l.h$ 。

最终的本地地形特征向量为 $tv(l)=\langle l.p_f, l.p_w, l.p_s, l.p_n, l.h \rangle$ 。其中, $l.h$ 采用 min-max 归一化进行处理,则本地地形特征向量的维度为 5。

2.2.3 k 近邻地形特征

待估计位置的空气质量与其周边区域的空气质量直接相关^[5],因此提取待估计位置周边空气质量监测站点的数据和地形作为 k 近邻地形特征。给定位置 l ,其 k 近邻地形特征的计算步骤如下:1)查询获取与 l 空间距离最近的 N 个空气质量监测站点,即 $l.m_1, \dots, l.m_N$;2)计算每个空气质量监测站点的地形特征向量,即 $tv(l.m_1), \dots, tv(l.m_N)$;3)计算 l 与 $l.m_1, \dots, l.m_N$ 间的空间距离,即 $sd(l, l.m_1), \dots, sd(l, l.m_N)$;4)基于欧氏距离计算 $tv(l)$ 与 $tv(l.m_1), \dots, tv(l.m_N)$ 间的向量距离,即 $vd(tv(l), tv(l.m_1)), \dots, vd(tv(l), tv(l.m_N))$ 。

最终的 k 近邻地形特征向量为 $k_v(l) = \langle sd(l, l.m_1), vd(tv(l), tv(l.m_1)), aqi(l.m_1), \dots, sd(l, l.m_N), vd(tv(l), tv(l.m_N)), aqi(l.m_N) \rangle$ 。其中, $aqi(l.m_N)$ 为空气质量监测站点 $l.m_N$ 当前的 AQI 真实值,则 k 近邻地形特征向量的维度为 $3N$ 。

综上,最终的原始特征向量为 $ov(l) = \langle mv(l), tv(l), k_v(l) \rangle$ 。表 1 列出了各类原始特征(给定当前位置 l)。

表 1 原始特征总结
Table 1 Summary of original features

特征类型	特征描述
气象特征	天气、温度、风力等级
时间特征	季节、是否为工作日、时间段
本地地形特征	l 周边森林地形占比、水域地形占比、沙漠地形占比、普通地形占比、 l 处海拔
k 近邻地形特征	第 1 个周边站点与 l 的空间距离、本地地形向量距离、真实 AQI、 $\dots$ 、第 k 个周边站点与 l 的空间距离、本地地形向量距离、真实 AQI

2.3 深层特征

在点击率预测^[11]领域的研究发现,对原始特征进行提炼可有效提高模型的性能,包括对数值型特征进行离散化、对类别型特征进行组合等。例如,海拔高到一定程度或低到一定程度才会对空气质量有明显影响,不同季节的白天和晚上的空气质量变化规律可能不同。然而,采用零散的特征提炼方法的效率很低。例如,数值型特征离散化存在如何确定离散边界的问题,类别型特征组合的复杂度太高且容易产生大量对提升模型性能毫无意义的组合。

决策树是一种对特征进行分段以分裂节点,对路径上的节点进行组合以实现估计的机器学习模型。其中,节点分裂可视为对数值型特征进行离散化,节点遍历可视为对类别型特征进行组合。因此,决策树可视为一种对特征进行提炼的整体解决方案。为保证特征提炼的多样性^[12],我们采用集成决策树模型 GBDT^[13]来生成深层特征(即经过提炼后的原始特征),具体步骤如下:

1)基于原始特征样本集训练一个 GBDT 模型。原始特征样本 $os(m, i) = (ov(m, i), aqi(m, i))$,其中 $ov(m, i)$ 和 $aqi(m, i)$ 分别为空气质量监测站点 m 在时刻 i 的原始特征向量和 AQI 真实值。

2)在 GBDT 模型的训练过程中,每次迭代都会根据上一次的残差生成一棵新的决策树。我们采用满二叉树结构建立决策树,并取出最后的 T 棵决策树用于生成深层特征。

3)对于原始特征样本 $os(l, i)$,将其依次输入每棵决策树, $os(l, i)$ 会根据其原始特征 $ov(l, i)$ 到达每棵决策树的某个叶子节点,则每个到达的叶子节点可看作一个深层特征,采用 one-hot 编码进行表示。如图 3 所示,假设取出了 3 棵深度为 3 的决策树,原始特征样本 $os(l, i)$ 到达了决策树 A 的叶子节点 2、决策树 B 的叶子节点 3 和决策树 C 的叶子节点 1,则采用 one-hot 编码得到的深层特征向量为 $\langle 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0 \rangle$ 。

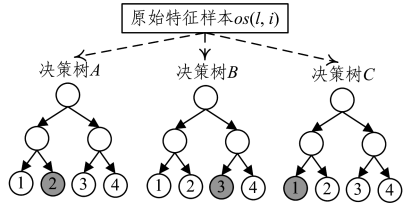


图 3 基于 GBDT 的深层特征的生成

Fig. 3 Generation of deep feature generation based on GBDT

给定位置 l ,最终的深层特征向量为 $dv(l) = \langle leaf_1(l), \dots, leaf_T(l) \rangle$ 。其中, $leaf_T(l)$ 为 l 对应的原始特征样本到达决策树 n 的叶子节点的编码。假设决策树深度为 H ,则深层特征向量的维度为 $2^{H-1} T$ 。综上,最终的特征向量为 $f_v(l) = \langle ov(l), dv(l) \rangle$,其维度为 $D = 21 + 3N + 2^{H-1} T$ 。

2.4 模型训练

假定特征向量 $f_v(x) = \langle x_1, x_2, \dots, x_D \rangle$,则回归函数如式(1)所示:

$$\hat{y} = w_0 + \sum_{i=1}^D w_i x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^D w_{ij} x_i x_j \tag{1}$$

其中, w_i 为特征分量 x_i 的回归系数, w_{ij} 为特征分量 x_i 和 x_j 组合的回归系数。在数据稀疏的情况下,难以保证特征分量 x_i 和 x_j 在训练样本中同时不为 0(特别是用 one-hot 编码的情况下),导致无法正常估计 w_{ij} 。

针对此问题,我们采用因子分解机^[14]作为回归模型。因子分解机基于矩阵分解的思想,将特征分量组合回归系数矩阵分解为两个低维矩阵的乘积(如式(2)所示)。其中, v_i 为一个 K 维向量($1 \leq i \leq D, K < D$),则 $W(i, j) = w_{ij} = v_i v_j^T = \sum_{k=1}^K v_{i,k} v_{j,k}$ 。利用矩阵分解的数据补全能力,可解决数据稀疏的问题。

$$W = VV^T = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{pmatrix} (v_1^T \quad v_2^T \quad \cdots \quad v_D^T) \tag{2}$$

定义了因子分解机模型后,需要从训练集中学习出参数 $\Theta = \{w_0, w_1, \dots, w_D, v_{1,1}, \dots, v_{1,K}, \dots, v_{D,1}, \dots, v_{D,K}\}$ 。类似于一般的有监督学习算法,因子分解机模型也可以通过定义和最优化损失函数来学习。如式(3)所示,目标函数由损失函数 L 和正则化项(用于防止过拟合)构成, S 为训练集合。我们采用 libFM 中集成的马尔科夫链蒙特卡罗算法(MCMC)对参数进行学习^[15]。

$$OPT(S) = \arg \min_{\Theta} \sum_{(x,y) \in S} L(y(x|\Theta), y) + \sum_{\theta \in \Theta} \lambda \theta^2 \tag{3}$$

3 实验

3.1 实验设置

1)实验数据。我们采集了从2016年1月1日至2016年12月31日全国范围内2266个空气质量监测站点的AQI数据,以及2469个地区的气象数据(包括天气、温度、风力等级)。其中,AQI每小时发布1次,气象数据每天发布2次。最终,数据集包含11576867条AQI数据和1698672条气象数据(空气质量和气象数据存在部分缺失的情况)。

我们从全国所有空气质量监测站点中挑选两组空气质量监测站点集(站点集A和站点集B)进行实验。其中,站点集A包含杭州市区范围内的10个空气质量监测站点,站点集B包含从全国范围内挑选的12个空气质量监测站点(包含森林地形面积占比最大的2个监测站点、水域地形面积占比最大的2个监测站点、海拔最高的2个监测站点,以及离上述每个监测站点空间距离最近的1个普通地形面积占比最大的监测站点)。站点集B的空气质量监测站点的详细信息如表2所列(其中,斜体站点编号代表该监测站点为自然地形站点,地形分布向量的4个分量分别为森林地形占比、水域地形占比、沙漠地形占比和普通地形占比,海拔的单位为m)。

表2 站点集B的详细信息

Table 2 Detailed information of station set B

站点编号	站点所在地区	地形分布向量	海拔/m
1266	三门峡市	$\langle 0, 0.98, 0, 0.02 \rangle$	332
1268	三门峡市	$\langle 0, 0.97, 0, 0.03 \rangle$	331
91956	运城市	$\langle 0, 0, 0, 1 \rangle$	369
91955	运城市	$\langle 0, 0, 0, 1 \rangle$	372
1333	宝鸡市	$\langle 0.98, 0, 0, 0.02 \rangle$	1041
1568	泉州市	$\langle 0.92, 0, 0, 0.08 \rangle$	245
92415	天水市	$\langle 0, 0, 0, 1 \rangle$	1504
1309	厦门市	$\langle 0, 0, 0, 1 \rangle$	213
92394	那曲地区市	$\langle 0, 0, 0, 1 \rangle$	4539
92397	阿里地区市	$\langle 0, 0, 0, 1 \rangle$	4347
92452	吐鲁番地区市	$\langle 0, 0, 0, 1 \rangle$	43
92304	乐山市	$\langle 0, 0.33, 0, 0.67 \rangle$	384

2)评测方式。针对某一站点集,本文采用“留一站点法”进行评测,即每次用其中一个站点的数据作为测试集,而其余站点的数据作为训练集。例如,站点集A的每个站点可产生 $24 \times 365 = 8760$ 个测试样本。

3)评测指标。本文采用均方根误差(RMSE)作为评测指标(见式(4))。其中, $y_i^{\wedge}$ 为真实值, y_i 为估计值。RMSE值越小,说明估计性能越好。

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

(4)

3.2 参数调试实验

本文提出的方法中,有4个关键参数需要调试,包括空气质量监测站点数量 N (见2.2.3节)、集成决策树的数量 T 和深度 H (见2.3节)、因子分解机的低维矩阵维度 K (见2.4节)。此外,正方形区域的边长 d (见2.2.2节)设置为6km。

实验1 参数 N 的调试

参数 N 用于生成 k 近邻地形特征。我们采用站点集A进行实验,待评测的 N 值包括 $\{0, 1, 2, 3, 4, 5, 6\}$ ($N=0$ 即不考虑 k 近邻地形特征)。此外,在该实验中,我们不考虑深层

特征,而参数 K 取默认值10,实验结果如图4所示。可以看出,首先, $N=0$ 时估计性能显著降低,证实了 k 近邻地形特征的有效性;其次,估计性能会随着 N 的增大而提升,直到 $N=4$,但继续增大 N 后,估计性能反而会出现少量降低。这说明某位置的空气质量与距离较近的空气质量监测站点的相关度较高,而与距离较远的空气质量监测站点的相关度较低,因此取 $N=4$ 。

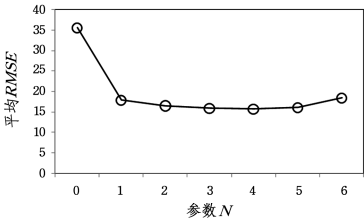


图4 参数 N 对估计性能的影响

Fig. 4 Influence of N on estimation performance

实验2 参数 T 和 H 的调试

参数 T 和 H 用于控制集成决策树的数量和深度,进而控制深层特征的维度。我们同样采用站点集A进行实验,待评测的 T 值包括 $\{10, 20, 30, 50, 100, 150, 200, 300\}$,待评测的 H 值包括 $\{3, 4, 5\}$,则待评测的参数组合有 $8 \times 3 = 24$ 个。在该实验中,参数 $N=4, K=10$,实验结果如图5所示。可以看出,首先,参数 T 过小(即决策树个数过少)时估计性能降低得较明显,而 T 增大到一定程度后对估计性能的影响明显降低;其次,参数 H (即决策树深度)对估计性能的总体影响较小。因此,我们取最优估计性能时的参数值,即 $T=50, H=4$ 。

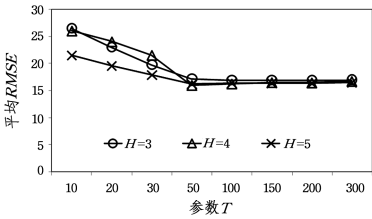


图5 参数 T 和 H 对估计性能的影响

Fig. 5 Influence of T and H on estimation performance

实验3 参数 K 的调试

参数 K 用于指定因子分解机的低维矩阵的维度。同样采用站点集A进行实验,待评测的 K 值包括 $\{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$ 。在该实验中,参数 $N=4, T=50, H=4$,则特征向量的维度为 $D = 21 + 3N + 2^{H-1}T = 433$,实验结果如图6所示。

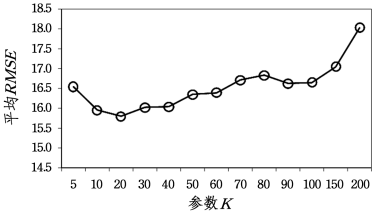


图6 参数 K 对估计性能的影响

Fig. 6 Influence of K on estimation performance

可以看出,首先,参数 K 对估计性能的总体影响不大;其次,当 $K>40$ 时,随着 K 的增大,因子分解机可用更高维度的

矩阵去补全特征分量组合回归系数矩阵,但是估计性能反而有一定程度的降低,这有可能是因为较小的 K 有助于避免过拟合,因此取 $K=20$ 。

3.3 特征有效性实验

实验 4 地形特征有效性的评测

为评测地形特征对空气质量的影响,对以下两组方法设置的估计性能进行比较。

1)设置 1(无地形特征):建立模型时不采用本地地形特征和 k 近邻地形特征中的向量距离(采用文本提出的其余特征建立模型)。

2)设置 2(有地形特征):采用本文提出的所有特征建立模型。

我们采用站点集 A 和站点集 B 进行实验,并评测地形特征对每个站点估计性能的影响。在该实验中,我们不考虑深层特征,而参数 $N=4, K=20$ 。站点集 A 的实验结果如表 3 所列,可以看出,由于站点集 A 是从城市市区范围内挑选的,因此地形特征对站点集 A 的估计性能的影响不大。站点集 B 的实验结果如表 4 所列,可以看出,地形特征对普通站点估计性能的影响不大,但对自然地形站点的估计性能有较大的影响。

表 3 站点集 A 的地形特征有效性评测

Table3 Validity evaluation of terrain features on station set A		
站点编号	无地形特征 $RMSE$	有地形特征 $RMSE$
1838	15.2125	15.3488
1839	15.5080	14.0157
1840	16.6343	16.7758
1841	14.5802	15.1208
1842	13.0190	13.2935
1843	16.2819	15.0141
1844	14.9424	13.3821
1845	20.6020	19.5791
1846	17.0129	17.1362
1847	20.1188	17.7845
平均	16.3912	15.7451

表 4 站点集 B 的地形特征有效性评测

Table 4 Validity evaluation of terrain features on station set B		
站点编号	无地形特征 $RMSE$	有地形特征 $RMSE$
1266	28.3226	27.8323
1268	26.6599	27.4246
91956	26.8297	30.5042
91955	32.1740	32.5110
1333	26.3737	26.0842
1568	15.3295	14.5885
92415	35.8199	32.6908
1309	25.7490	30.0328
92394	90.7257	86.8605
92397	91.3195	63.7485
92452	113.910	115.936
92304	13.1297	13.1135
平均	43.8619	41.7772
普通地形站点平均	41.2687	42.4647
自然地形站点平均	46.4552	41.0898

实验 5 深层特征有效性评测

为评测深层特征的效果,对以下两组方法设置的估计性能进行比较。

1)设置 1(无深层特征):采用本文提出的原始特征建立模型。

2)设置 2(有深层特征):采用本文提出的所有特征建立模型。

我们同样采用站点集 A 和站点集 B 进行实验。在该实验中,参数 $N=4, T=50, H=4, K=20$ 。站点集 A 和站点集 B 的实验结果分别如表 5 和表 6 所列。可以看出,加入深层特征后,无论是对普通地形区域还是自然地形区域的监测站点,其估计性能总体都有所提升。

表 5 站点集 A 的深层特征有效性评测

Table 5 Validity evaluation of deep features on station set A		
站点编号	无深层特征 $RMSE$	有深层特征 $RMSE$
1838	15.3488	14.0774
1839	14.0157	14.3538
1840	16.7758	16.7689
1841	15.1208	14.5934
1842	13.2935	12.9171
1843	15.0141	14.4831
1844	13.3821	13.2036
1845	19.5791	18.6106
1846	17.1362	16.5092
1847	17.7845	14.4665
平均	15.7451	14.9984

表 6 站点集 B 的深层特征有效性评测

Table 6 Validity evaluation of deep features on station set B		
站点编号	无深层特征 $RMSE$	有深层特征 $RMSE$
1266	27.8323	26.0437
1268	27.4246	26.6135
91956	30.5042	27.4629
91955	32.5110	32.2340
1333	26.0842	23.8250
1568	14.5885	13.2587
92415	32.6908	26.2067
1309	30.0328	25.3810
92394	86.8605	85.6407
92397	63.7485	43.2915
92452	115.936	114.859
92304	13.1135	13.1827
平均	41.7772	38.1666
普通地形站点平均	42.4647	39.8877
自然地形站点平均	41.0898	36.4455

3.4 对比实验

实验 6 与基线方法的对比

我们对以下 5 种 AQI 空间估计方法的性能进行了对比。

1)LI:线性插值方法,以与周边空气质量监测站点的空间距离为权重进行插值(见式(5))。

$$aqi(l)=\frac{\sum_{i=1}^N aqi(l,m_i)\times \frac{1}{sd(l,l,m_i)}}{\sum_{i=1}^N \frac{1}{sd(l,l,m_i)}}\tag{5}$$

其中, $sd(l,l,m_i)$ 表示当前位置 l 与周边第 i 个站点的空间距离。

2)LR:土地利用回归方法,采用线性回归算法建立空气质量空间估计模型,其中特征采用的是本文提出的原始特征(参数 $N=4$)。

3)GBDT:采用 GBDT 算法建立空气质量空间估计模型,其中特征采用的是本文提出的原始特征(参数 $N=4, T=50, H=4$)。

4)FM:采用因子分解机算法建立空气质量空间估计模

型,其中特征采用的是本文提出的原始特征(参数 $N=4, K=20$)。

5)OUR:本文提出的方法(参数 $N=4, T=50, H=4, K=20$)。

采用站点集 A 和站点集 B 进行实验,实验结果如图 7 所示。可以看出:

1)无论对普通地形区域还是对自然地形区域的监测站点,本文方法的估计性能均最优。

2)对于站点集 A,简单的线性插值方法 LI 的估计性能与本文方法 OUR 的差别不大,甚至略优于部分机器学习方法(包括 LR 和 GBDT)。这说明在不引入更丰富因素(如交通、人流等)的情况下,使用机器学习方法估计普通地形区域的 AQI 收益并不理想。

3)对于站点集 B,机器学习方法(包括 GBDT,FM,OUR)的估计性能明显优于线性插值方法 LI。这说明在自然地形区域中,地形特征对空气质量有显著的影响,而机器学习方法可有效地学习到这种影响。

4)FM 的估计性能始终优于 GBDT。这证实了基于特征分量组合训练模型和基于矩阵分解处理数据稀疏的策略在 AQI 空间估计问题中的有效性。

5)LR 的估计性能始终较差。这说明空气质量随位置的变化是非线性的,而线性模型难以学习到这种规律。

6)对站点集 B 的估计性能明显低于对站点集 A 的估计性能。可能有如下原因:站点集 B 中的站点是从全国范围内挑选的,而空间距离过大的站点(如福建的站点和新疆的站点),即使地形分布相似,空气质量的变化规律也可能相差很大。

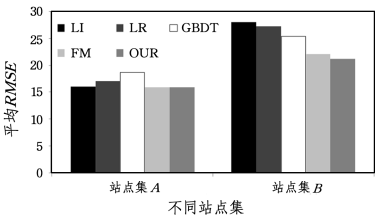


图 7 不同方法的估计性能比较

Fig. 7 Estimation performance comparison of different methods

结束语 为估计无空气质量监测站点位置的 AQI,本文提出了一种基于地形特征的 AQI 空间估计方法。本文方法首先基于地形数据建立原始特征,然后基于集成决策树模型对原始特征进行深层转换,最后基于因子分解机建立回归模型。通过大量实验发现,本文方法对估计自然地形区域中的 AQI 有着明显的优势,可有效弥补现有方法只适应普通地形区域的不足。

今后可从如下几个方面进行深入研究:1)绝大部分空气质量监测站点都部署在普通地形区域,因此某一地区可能会缺少甚至没有部署在自然地形区域的空气质量监测站点,可探索采样、半监督学习等手段来解决这一问题;2)各种因素对空气质量的影响在不同地区可能有所不同,因此在一个地区建立的模型在另一个地区的估计性能可能较差,可探索迁移学习等手段来解决这一问题。

参 考 文 献

[1] HUANG J,ZHOU C,LEE X,et al. The effects of rapid urbanization on the levels in tropospheric nitrogen dioxide and ozone over East China [J]. Atmospheric Environment, 2013, 77(1): 558-567.

[2] WATSON D F,PHILIP G M. A refinement of inverse distance weighted interpolation [J]. Geoprocessing, 1985, 2(4): 315-327.

[3] SHAD R,MESGARI M S,ABKAR A,et al. Predicting air pollution using fuzzy genetic linear membership kriging in GIS [J]. Computers, Environment and Urban Systems, 2009, 33(6): 472-481.

[4] ZHENG Y,LIU F,HSIEH H P. U-Air: When urban air quality inference meets big data [C]// Proceedings of ACM Conference on Knowledge Discovery and Data Mining. 2013: 1436-1444.

[5] CHEN L,CAI Y,DING Y,et al. Spatially fine-grained urban air quality estimation using ensemble semi-supervised learning and pruning [C]// Proceedings of the ACM Joint Conference on Pervasive and Ubiquitous Computing. 2016: 1076-1087.

[6] HOEK G,BEELEN R,DE HOOGH D,et al. A review of land-use regression models to assess spatial variation of outdoor air pollution [J]. Atmospheric Environment, 2008, 42(1): 7561-7578.

[7] HO C C,CHAN C C,CHO C W,et al. Land use regression modeling with vertical distribution measurements for fine particulate matter and elements in an urban area [J]. Atmospheric Environment, 2015, 104(1): 256-263.

[8] JUTZELER A,LI J J,FALTINGS B. A region-based model for estimating urban air pollution [C]// Proceedings of AAAI Conference on Artificial Intelligence. 2014: 424-430.

[9] HASENFRATZ D,SAUKH O,WALSER C,et al. Pushing the spatio-temporal resolution limit of urban air pollution maps [C]// Proceedings of IEEE Conference on Pervasive Computing and Communications. 2014: 69-77.

[10] KIM D,STOCKWELL W R. An online coupled meteorological and air quality modeling study of the effect of complex terrain on the regional transport and transformation of air pollutants over the Western United States [J]. Atmospheric Environment, 2008, 42(17): 4006-4021.

[11] HE X,PAN J,JIN O,et al. Practical lessons from predicting clicks on ads at Facebook [C]// Proceedings of ACM Conference on Knowledge Discovery and Data Mining. 2014: 1-9.

[12] LI N,YU Y,ZHOU Z H. Diversity regularized ensemble pruning [C]// Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2012: 330-345.

[13] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine [J]. Annals of Statistics, 1999, 29(1): 1189-1232.

[14] RENDLE S. Factorization machines [C]// Proceedings of IEEE International Conference on Data Mining. 2010: 995-1000.

[15] RENDLE S. Factorization machines with libFM [J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(3): 1-22.