工程科学学报
Chinese Journal of Engineering

基于图像混合核的列生成$PM_{2.5}$预测

李晓理 张博 杨旭

Column-generation $PM_{2.5}$ prediction based on image mixture kernel

LI Xiao-li, ZHANG Bo, YANG Xu

引用本文:
李晓理, 张博, 杨旭. 基于图像混合核的列生成$PM_{2.5}$预测[J]. 工程科学学报, 2020, 42(7): 922-929. doi: 10.13374/j.issn2095-9389.2019.07.15.002
LI Xiao-li, ZHANG Bo, YANG Xu. Column-generation $PM_{2.5}$ prediction based on image mixture kernel[J]. *Chinese Journal of Engineering*, 2020, 42(7): 922-929. doi: 10.13374/j.issn2095-9389.2019.07.15.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2019.07.15.002

工程科学学报, 第 42 卷, 第 7 期: 922−929, 2020 年 7 月
Chinese Journal of Engineering, Vol. 42, No. 7: 922−929, July 2020
https://doi.org/10.13374/j.issn2095-9389.2019.07.15.002; http://cje.ustb.edu.cn

基于图像混合核的列生成 $PM_{2.5}$ 预测

李晓理[1,2,3,4)], 张　博[1)], 杨　旭[5)]✉

1) 北京工业大学信息学部, 北京 100124　2) 计算智能与智能系统北京市重点实验室, 北京 100124　3) 数字社区教育部工程研究中心, 北京 100124　4) 北京未来网络科技高精尖创新中心, 北京 100124　5) 北京科技大学自动化学院, 北京 100083
✉通信作者, E-mail: yangxu@ustb.edu.cn

摘　要　传统 $PM_{2.5}$ 预测方法获取污染物浓度数据需要大型精密仪器, 成本较高. 本文尝试利用图像数据进行 $PM_{2.5}$ 浓度预测. 大气 $PM_{2.5}$ 浓度的变化与图像的暗通道强度、对比度和 HSI(Hue-saturation-intensity)颜色差异有密切联系. 大气中 $PM_{2.5}$ 浓度的升高会导致非天空区域的暗通道强度值下降, 图像对比度下降和 HSI 空间颜色差异变小. 通过分析 $PM_{2.5}$ 浓度与图像特征的关系, 提出了一种基于图像混合核的列生成空气质量 $PM_{2.5}$ 预测模型. 首先, 以 1 h 为采样周期, 每日 8:00 ~ 17:00 为采样范围, 采集多种天气条件下的景物图像, 提取图像的对比度、暗通道强度和 HSI 颜色差异共 5 个图像特征. 其次, 数据存在样本规模大、样本不平坦分布等特点, 单个核函数构成的预测模型难以满足预测精度需求, 因此本文按照核结构从简单到复杂的原则, 选择线性核函数、多项式核函数和高斯核函数三种核函数建立组合模型. 然后计算每个核基于训练样本的 Gram 矩阵, 并将所有 Gram 矩阵并列成一个混合核矩阵. 利用列生成算法和混合核矩阵建立预测模型, 求解模型参数. 最后, 进行仿真实验, 实验结果表明本文提出的可满足预测精度要求, 与单核预测模型相比, 该预测模型预测精度更高, 模型稳定性更好. 计算复杂度分析结果显示基于图像混合核的列生成模型与单核预测模型相比计算量无明显增加.

关键词　$PM_{2.5}$ 预测; 混合核函数; 列生成算法; 图像特征; 预测模型
分类号　TP181

Column-generation $PM_{2.5}$ prediction based on image mixture kernel

LI Xiao-li[1,2,3,4)], *ZHANG Bo*[1)], *YANG Xu*[5)]✉

1) Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
2) Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China
3) Engineering Research Center of Digital Community, Beijing 100124, China
4) Beijing Future Network Science and Technology Innovation Center, Beijing 100124, China
5) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
✉ Corresponding author, E-mail: yangxu@ustb.edu.cn

ABSTRACT　The conventional method of $PM_{2.5}$ prediction requires high-precision instruments to obtain data on the concentration of pollutants, resulting in a high prediction costs. In this work, we attempt to use image data to estimate $PM_{2.5}$ concentration. The concentration of atmospheric $PM_{2.5}$ is closely linked to the image's dark channel intensity, contrast, and color difference of HSI. The increase in atmospheric $PM_{2.5}$ concentration leads to a decrease in the non-sky area dark channel intensity, image contrast, and HSI spatial color difference. In this paper, a Column-Generation $PM_{2.5}$ prediction model based on image mixture kernel was proposed by analyzing the relationship between $PM_{2.5}$ and image features. First, the sampling period was taken as 1 h, and 8:00–17:00 was taken as the sampling range daily. The scene images were recorded in different weather conditions, and five image features were extracted,

收稿日期: 2019−07−15
基金项目: 国家自然科学基金资助项目(61873006, 61473034, 61673053); 北京市科学重大专项资助项目(Z181100003118012); 国家重点研发计划资助项目(2018YFC1602704, 2018YFB1702704)

including contrast, dark channel intensity, and HSI color difference. Secondly, the image data has the characteristics of large sample size and uneven distribution, and the prediction model consists of a single kernel function, which makes it difficult to meet the prediction accuracy requirement. Therefore, the linear kernel function, polynomial kernel function, and Gauss kernel function were chosen to construct a composite model according to the concept of kernel structure from simple to complex. Then each kernel's Gram matrix was calculated based on training samples, and all gram matrices were placed into a mixture kernel matrix. Using the column generation algorithm and mixture kernel matrix, the prediction model was developed and the parameters of the model were solved. Finally, simulation experiments were performed; the results show that the prediction model based on the image mixture kernel of Column-Generation $PM_{2.5}$ can meet the prediction accuracy requirements. The model has higher prediction accuracy and better model stability in comparison with the single-kernel prediction model. A computational complexity analysis shows that the prediction model based on the image mixture kernel of column-generation $PM_{2.5}$ has no significant increase in computational complexity in comparison with the one-kernel prediction model.

KEY WORDS $PM_{2.5}$ prediction; mixed kernel function; column generation algorithm; image feature; prediction model

伴随着雾霾在全国各地出现的频率越来越高，环境问题越来越引起人们的关注，尤其以京津冀地区最为明显．$PM_{2.5}$ 是造成雾霾的主要因素，其在空气中滞留时间长，通过对太阳光的吸收、散射或反射，降低环境可见度；$PM_{2.5}$ 颗粒被吸入人体后，会渗透到肺部组织，引发支气管炎等疾病，对人体健康造成危害．加强大气环境污染控制已成为亟待解决的问题，通过对 $PM_{2.5}$ 预测，可以为环境治理和人们健康出行提供准确的环境质量信息，有助于减轻环境污染对群众造成的危害．

近年来，国内外学者都对 $PM_{2.5}$ 预测方法进行了卓有成效的研究．文献 [1] 基于单时间序列模型，将动态指数平滑法和动态马尔科夫模型相结合，通过 $PM_{2.5}$ 历史数据预先确定算法的最优参数，对 $PM_{2.5}$ 进行动态预测，并验证了模型的有效性．文献 [2] 通过构建空间平滑核，对梯度增强算法进行改进，有效解决了 $PM_{2.5}$ 浓度与气溶胶光学深度、气象条件等预测变量之间的空间非平稳性，对日 $PM_{2.5}$ 进行预测．文献 [3] 提出了一种基于主成分分析和最小二乘支持向量机的杜鹃搜索混合模型，并将模型的预测效果与广义回归神经网络模型作对比，效果更优．文献 [4] 提取大气温度、湿度和风速 3 个特征，训练长短期记忆模型对 1 h 后的 $PM_{2.5}$ 污染等级进行预测，证明了 $PM_{2.5}$ 污染与周边地区的气象条件有密切联系．文献 [5] 利用随机数据分析方法，在多变量系统中选择与 $PM_{2.5}$ 相关的随机变量，作为神经网络的输入，实现了空气质量预测．文献 [6] 建立基于互补集合经验模态分解和支持向量回归的混合预测模型．对 $PM_{2.5}$ 质量浓度的原始时间序列进行分解，得到若干具有不同时间尺度的相对平稳分量，采用 SVR 算法对各个分量分别进行预测，求出各个分量的预测值之和，作为原始 $PM_{2.5}$ 质量浓度的预测结果．

也有学者利用图像对 $PM_{2.5}$ 进行预测．文献 [7] 利用大量室外图像，结合太阳位置、日期、时间、地理信息、天气条件等相关数据对 $PM_{2.5}$ 进行预测，该方法避免了大气测量装置的限制，为预测 $PM_{2.5}$ 提供了一种更为便捷的方式．文献 [8] 以手机照片为数据源，对良好天气下空间域和变换域的图像熵值建立自然度统计模型，通过计算污染图像的熵值的偏差度对 $PM_{2.5}$ 进行预测．文献 [9] 对不同天气条件下的照片质量进行分析建模，通过提取 $PM_{2.5}$ 浓度相关的特征建立粒子群优化的支持向量回归模型，实现了良好的预测效果．

鉴于大气环境复杂多变，$PM_{2.5}$ 预测中需要考虑的因素较多，本文在上述研究的基础上，提出了一种基于图像混合核的列生成 $PM_{2.5}$ 预测方法．该方法通过分析图像变化与 $PM_{2.5}$ 浓度的关系，提取图像特征，并利用相关性分析完成特征选择．将图像特征经混合核映射到高维线性空间，有效避免了单核函数选取不当造成的影响．最后使用列生成方法来求解模型参数，保证了解的稀疏性和精确性，实现了对 $PM_{2.5}$ 的有效预测．

1 问题描述与数学基础知识

1.1 问题描述

$PM_{2.5}$ 指空气动力学当量直径小于等于 2.5 μm 的悬浮物颗粒，会对可见光产生明显的散射作用．大气中 $PM_{2.5}$ 浓度的变化会使图像呈现不同的对比度、暗通道强度、可见度等特征信息，这使得利用图像实现 $PM_{2.5}$ 预测成为可能．

1.2 数学基础知识

本文采用了基于图像混合核的列生成方法研究了 $PM_{2.5}$ 预测问题，为了更好地介绍理论的原

理, 下面对方法中需要用到的一些数学基础知识进行简单地说明.

1.2.1 混合核

核方法被证明了是解决许多应用中推理问题的有效方法. 通过引入正半定核$\boldsymbol{K}$, 可以使用线性学习算法创建非线性模型. 给定观测样本$\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}\in\boldsymbol{X}\times\boldsymbol{Y}$. 其中输入空间$\boldsymbol{X}\in\mathbf{R}^n$, 输出空间$\boldsymbol{Y}\in\mathbf{R}$(回归问题), 通过非线性映射:

$$\begin{aligned}&\boldsymbol{\Phi}:\boldsymbol{X}\to\boldsymbol{F}\\&\boldsymbol{x}\mapsto\boldsymbol{\Phi}(\boldsymbol{x})\end{aligned}\tag{1}$$

把输入数据映射到一个新的特征空间$\boldsymbol{F}=\{\boldsymbol{\Phi}(\boldsymbol{x})|\boldsymbol{x}\in\boldsymbol{X}\}$, 其中$\boldsymbol{F}\in\mathbf{R}^n$, 原问题转化为:

$$\{(\boldsymbol{\Phi}(\boldsymbol{x}_1),y_1),(\boldsymbol{\Phi}(\boldsymbol{x}_2),y_2),\cdots,(\boldsymbol{\Phi}(\boldsymbol{x}_l),y_l)\}\in\boldsymbol{F}\times\boldsymbol{Y}\tag{2}$$

在满足 Mercer 条件情况下, 一定存在一个特征空间$\boldsymbol{F}$和一个映射$\boldsymbol{\Phi}:\boldsymbol{X}\to\boldsymbol{F}$, 使得

$$k(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{\Phi}(\boldsymbol{x})\times\boldsymbol{\Phi}(\boldsymbol{z})\tag{3}$$

$k(\boldsymbol{x},\boldsymbol{z})$即为核函数.

核函数有两种主要的类型: 全局核函数和局部核函数, 局部性核函数学习能力强、泛化性能较弱, 而全局性核函数泛化性能强、学习能力较弱, 因此考虑把这两类核函数混合起来构成混合核函数. 对文献 [10] 中混合核函数的形式进行扩展得到多核混合核函数的形式为$k(\boldsymbol{x},\boldsymbol{z})=\sum_{p=1}^{P}\mu_p k_p(\boldsymbol{x},\boldsymbol{z})$, 其中$k_p(\boldsymbol{x},\boldsymbol{z})$为单核函数, p是对应的核函数编号, μ_p为组合系数. 由 SVM 决策函数可知, 混合核函数的决策函数为:

$$f(x)=\sum_{j=1}^{l}\alpha_j\left(\sum\nolimits_p\mu_p k_p(\boldsymbol{x},\boldsymbol{x}_j)\right)\tag{4}$$

式中, $\boldsymbol{\alpha}$是模型参数, $\boldsymbol{x}_j$是第j个输入向量. 本文中, 不单独计算每个核矩阵(核对样本的 Gram 矩阵), 而是采用混合模型, 其决策函数为:

$$f(x)=\sum_{j=1}^{l}\sum_{p=1}^{P}\alpha_j^p k_p(\boldsymbol{x},\boldsymbol{x}_j)\tag{5}$$

1.2.2 列生成

列生成算法是用于求解大型线性规划问题的一种重要方法. 在原始问题中, 列生成算法并不是一次性求解出所有参数$\boldsymbol{\alpha}$, 而是选取混合核矩阵$\boldsymbol{K}$(构造方法在第 4 章介绍)的列子集并求解对应的$\boldsymbol{\alpha}$的最优解[11]. 根据拉格朗日对偶性[12], 通过求解对偶问题可得到原始问题的最优解. 原始问题的每一列对应于对偶问题的一个约束, 当约束问题的解违反对偶问题中不存在的约束时, 则需将该约束(原始问题中的一列)添加到约束问题中, 以获得最优解.

基于决策函数(5), 重写文献 [13] 中的线性列生成增强算法, 使用 2 范数正则化构建如下凸二次规划问题:

$$\begin{aligned}&\min\nolimits_{\alpha,\xi}\frac{1}{2}\sum_{j=1}^{d}\alpha_j^2+C\sum_{i=1}^{l}\xi_i\\&\text{s.t.}\quad y_i\sum_{j=1}^{d}\boldsymbol{K}_{ij}\alpha_j+\xi_i\geqslant 1,\xi_i\geqslant 0,i=1,\cdots,l,\\&\alpha_i\geqslant 0,j=1,\cdots,d\end{aligned}\tag{6}$$

求得其对偶问题为:

$$\begin{aligned}&\max\nolimits_u\min\nolimits_\alpha\sum_{i=1}^{l}u_i-\frac{1}{2}\sum_{j=1}^{d}\alpha_j^2\\&\text{s.t.}\quad\sum_{i=1}^{l}u_i y_i\boldsymbol{K}_{ij}\leqslant\alpha_j,j=1,\cdots,d,\\&0\leqslant u_i\leqslant C,i=1,\cdots,l\end{aligned}\tag{7}$$

求解式(6)和(7)的最优解为$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{u}})$, 根据文献 [13], 验证如下问题:

$$\tau=\max_j\sum_{i=1}^{l}\hat{u}_i y_i\boldsymbol{K}_{ij}\tag{8}$$

式中, j遍历核矩阵$\boldsymbol{K}$中的所有列. 列生成算法将列系数$\boldsymbol{\alpha}$分为两部分, 使用启发式算法选出的一部分W用于训练模型, 未选中的部分N作为备选, 假设未选中的部分$\boldsymbol{\alpha}^N=0$, 通过求解式(6)和(7)得当前最优解得$\boldsymbol{\alpha}^W$, 则$\hat{\boldsymbol{\alpha}}=(\boldsymbol{\alpha}^W,\boldsymbol{\alpha}^N=0)$. 经文献 [14] 证明, $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{u}})$是原始−对偶问题的当前最优解, 如果对于所有的$j\in N,\sum_{i=1}^{l}u_i y_i\boldsymbol{K}_{ij}\leqslant 0$, 则$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{u}})$即为满足 KKT 条件的全局最优解. 对于线性列生成增强模型, 每次选择N中使$\sum_{i=1}^{l}u_i y_i\boldsymbol{K}_{ij}$最大的列$K_{\cdot j}$加入到约束问题中.

将列生成增强算法推广到解决具有不敏感参数ε的损失函数$\max\{|y-f(x)|-\varepsilon,0\}$的回归问题[15], 模型的下限约束$\alpha>0$为非必需条件, 所以在原模型中去除下限约束. 为了构建回归模型, 本文将偏离真实值至少ε的点作为误差点. 使用 2 范数正则化, 对应的凸二次规划问题为:

$$\begin{aligned}&\min\nolimits_{\alpha,\xi,\eta}\frac{1}{2}\sum_{j=1}^{d}\alpha_j^2+C\sum_{i=1}^{l}(\xi_i+\eta_i)\\&\text{s.t.}\quad\sum_{i=1}^{l}\boldsymbol{K}_{ij}\alpha_j+\xi_i\geqslant y_i-\varepsilon,i=1,\cdots,l,\\&-\sum_{i=1}^{l}\boldsymbol{K}_{ij}\alpha_j+\eta_i\geqslant -y_i-\varepsilon,i=1,\cdots,l,\\&\xi_i\geqslant 0,\eta_i\geqslant 0,i=1,\cdots,l.\end{aligned}\tag{9}$$

设u_i, v_i为拉格朗日乘子，则原始问题(9)的对偶问题为：

$$\begin{aligned}&\max_{u,v}\min_{\alpha}\frac{1}{2}\sum_{j=1}^{d}\alpha_j^2+\sum_{i=1}^{l}(u_i-v_i)y_i-\sum_{i=1}^{l}(u_i+v_i)\varepsilon\\&\text{s.t.}\quad\sum_{i=1}^{l}(u_i-v_i)\boldsymbol{K}_{ij}=\alpha_j\end{aligned}\tag{10}$$

同理，求解如下问题：

$$\tau=\max_{j\in N}\left|\sum_{i=1}^{l}(\hat{u}_i-\hat{v}_i)\boldsymbol{K}_{ij}\right|\tag{11}$$

解为$\boldsymbol{K}_{\cdot\hat{j}}$. 经文献 [14] 证明，若$\tau=0$，则当前最优解$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$即为回归问题的全局最优解，否则，将$\boldsymbol{K}_{\cdot\hat{j}}$加入到约束问题中去.

2 图像特征提取与相关性分析

空气中的雾霾会对图像造成严重的影响，会导致图像的一些特征值变低，尤其会影响图像的对比度、视见度、暗通道强度等[16]. 本文提取多个与雾霾相关的图像特征，并将图像特征与 $PM_{2.5}$ 值做相关性分析完成特征选择.

2.1 特征提取

本节提取与 $PM_{2.5}$ 浓度相关的空间对比度、非天空区域的暗通道强度、HSI 空间颜色差异等特征.

2.1.1 空间对比度(F_{ig})

大气透射是指光线从场景辐射到观察者时，减去空气中颗粒物等的折射剩余的部分，是一个 0 到 1 之间的标量. 根据大气透射模型，大气光的消光与透射率呈反比关系，两者满足如下公式[17]：

$$t(x)=\exp^{-b_{\text{ext}}r(x)}\tag{12}$$

式中，b_{ext}是消光系数，$r(x)$是光的传输距离. 根据文献 [18]：

$$|\nabla_x I(x)|=t(x)|\nabla_x J(x)|\tag{13}$$

定义空间对比度F_{ig}为：$F_{\text{ig}}=|\nabla_x I(x)|$.

2.1.2 暗通道强度(F_{id})

图像的暗通道强度定义为[19]：

$$J_{\text{dark}}(x)=\min_{y\in\Omega(x)}\left\{\min_{c\in\{r,g,b\}}J^c(y)\right\}\tag{14}$$

式中，$\Omega(x)$是以像素x为中心的分块，J为场景辐射光，J^c表示其中一个颜色通道. 从式中可以看出，给定像素的暗通道强度值为该分块三颜色同道中的最小值. 大量无雾霾图像的先验知识表明，无雾霾图像的暗通道强度值为 0，即：

$$J_{\text{dark}}\to 0\tag{15}$$

将式(14)和(15)代入大气透射模型中，得：

$$t(x)=1-\min_{y\in\Omega(x)}\left\{\min_{c}\frac{I^c(y)}{A^c}\right\}\tag{16}$$

式中，A^c为大气光，因此将$t(x)$选为特征F_{id}.

2.1.3 HSI 颜色差异(F_{ih}, F_{is}, F_{ii})

根据 Kim 等的研究[20]，天空在 HSI 颜色空间中颜色差异与大气消光b_{ext}存在指数关系，可表示为：$b_{\text{ext}}=ae^{b\Delta D}$，式中$a$和$b$为模型参数，$\Delta D$用来描述 HSI 空间中的差异. 由于很难获取$b_{\text{ext}}$中在 HSI 三部分的影响参数，因此使用三部分在 HSI 颜色空间的差异值作为特征，定义如下：

$$\begin{aligned}F_{\text{ih}}&=\frac{1}{m*n}\sum_{y=1}^{n}\sum_{x=1}^{m}\sqrt{d_{\text{h}}(x)^2+d_{\text{h}}(\text{y})^2}\\d_{\text{h}}(x)&=I_{\text{h}}(x,y)-I_{\text{h}}(x+1,y)\\d_{\text{h}}(y)&=I_{\text{h}}(x,y)-I_{\text{h}}(x,y+1)\end{aligned}\tag{17}$$

式中，I是输入图像，其像素为$m*n$，$I_{\text{h}}(x,y)$是像素点(x,y)的h值. 同样，F_{is}和F_{ii}定义如下：

$$F_{\text{is}}=\frac{1}{m*n}\sum_{y=1}^{n}\sum_{x=1}^{m}\sqrt{d_{\text{s}}(x)^2+d_{\text{s}}(y)^2}\tag{18}$$

$$F_{\text{ii}}=\frac{1}{m*n}\sum_{y=1}^{n}\sum_{x=1}^{m}\sqrt{d_{\text{i}}(x)^2+d_{\text{i}}(y)^2}\tag{19}$$

2.2 相关性分析

采用皮尔逊相关系数对图像特征进行相关性计算. 皮尔逊相关系数广泛用于度量两个变量之间的相关程度，其值介于–1 与 1 之间，其中 1 表示完全正相关. 其形式如下：

$$r=\frac{\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^2}\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}})^2}}\tag{20}$$

特征与 $PM_{2.5}$ 浓度值相关性越强，该特征越能表示图像的信息. 当两组数据相关性系数大于 0.6 时，可认为两组数据相关性较强，当相关性系数小于 0.6 时认为两组数据相关性较弱. 因此，本文选择与 $PM_{2.5}$ 相关性系数绝对值大于 0.6 的特征作为最终模型训练特征.

3 基于图像混合核的列生成 $PM_{2.5}$ 预测

$PM_{2.5}$ 浓度变化主要影响图像对比度、非天空区域的暗通道强度、HSI 空间颜色差异. 由于图像特征与 $PM_{2.5}$ 浓度呈非线性关系[21]，考虑到普通核函数各有利弊，为了得到学习能力和泛化能力都很强的核函数，采用混合核的方法建立图像特征

值与 $PM_{2.5}$ 浓度之间的关系方程, 并利用列生成算法求解方程参数.

3.1　特征选择

从图像中提取 F_{ih}, F_{is}, F_{ii}, F_{ig}, F_{id} 共 5 个特征, 对各特征和 1 h 后的 $PM_{2.5}$ 值进行相关性分析, 结果如表 1 所示. 5 个特征与 $PM_{2.5}$ 浓度均呈负相关, 可知 $PM_{2.5}$ 浓度升高, 会导致图像对比度、暗通道强度下降, HSI 颜色差异变小. 其中 F_{ig},F_{id} 与 $PM_{2.5}$ 值的相关性强, F_{ih}, F_{is}, F_{ii} 与 $PM_{2.5}$ 的相关性较强. 因此, 本文选择 F_{ih}, F_{is}, F_{ii}, F_{ig}, F_{id} 共 5 个特征进行模型训练.

表 1　特征与 $PM_{2.5}$ 相关性值

Table 1　Correlation between characteristics and $PM_{2.5}$

F_{ig}	F_{id}	F_{ih}	F_{is}	F_{ii}
− 0.55	− 0.46	− 0.36	− 0.4	− 0.29

3.2　预测步骤

为方便预测, 首先构造混合核矩阵. 将给定的多个核函数组成核函数集 $\boldsymbol{S}=\{\boldsymbol{K}_1,\boldsymbol{K}_2,\cdots,\boldsymbol{K}_p\}$, 计算每个核基于训练样本 $\boldsymbol{K}_p(\cdot,x_j)$ 的 Gram 矩阵 $\boldsymbol{K}^p$ ($\boldsymbol{K}_p(\cdot,x_j)$ 对应第 j 个训练样本). 然后, 将所有 Gram 矩阵并列构成一个混合核矩阵 $\boldsymbol{K}=[\boldsymbol{K}^1,\boldsymbol{K}^2,\cdots,\boldsymbol{K}^p]$, 则 $\boldsymbol{K}$ 为 $l\times d$ 的矩阵, 其中 $d=l\times p$, $\boldsymbol{K}_{i\cdot}$ 表示混合核矩阵的第 i 行, $\boldsymbol{K}_{\cdot j}$ 表示混合核矩阵的第 j 列.

在没有任何先验知识的前提下, 优先选择简单的、计算成本低的核函数. 本实验中, 当简单核函数对应的列没有可添加的列用于求解时, 则需要从更加复杂的核函数列中选取列用于求解. 因此实验从简单到复杂采用三种核函数: 线性核函数 (L)、多项式核函数 (P)、RBF 核函数 (R) 构建混合核. 将给定的 3 个核函数组成核函数集 $\boldsymbol{S}=\{\boldsymbol{K}_L,\boldsymbol{K}_P,\cdots,\boldsymbol{K}_R\}$, 分别计算每个核基于训练样本的 Gram 矩阵 $\boldsymbol{K}^L,\boldsymbol{K}^P,\boldsymbol{K}^R$. 将所有 Gram 矩阵并列构成一个混合核矩阵 $\boldsymbol{K}=[\boldsymbol{K}^L,\boldsymbol{K}^P,\boldsymbol{K}^R]$, 然后基于混合核矩阵利用列生成算法求解模型参数. 实验中, L,P,R 表示单核预测模型, $L+P+R$ 表示本文提出的混合核模型, 核函数中的标准差 σ 用 $\|x_i-x_j\|^2$ 的均值代替 (i,j 遍历所有的训练样本). 基于图像混合核的列生成预测步骤如下.

步骤 1: 采集图像数据和 $PM_{2.5}$ 浓度数据, 经数据预处理后, 配成样本对;

步骤 2: 提取图像特征, 与 1 h 后的 $PM_{2.5}$ 浓度数据做相关性分析, 剔除弱相关特征;

步骤 3: 选取多个核函数, 计算核函数基于图像特征值的 Gram 矩阵;

步骤 4: 将多个 Gram 矩阵合并为混合核矩阵;

步骤 5: 抽取混合核矩阵的部分列构成列子集, 利用列生成算法基于列子集求取模型当前解;

步骤 6: 验证当前解是否为最优解. 若是, 输出最优解, 模型构建完成; 若否, 抽取未选列中的最佳列添加到列子集中, 返回步骤 5;

步骤 7: 利用验证集验证预测模型的精度与稳定性.

3.3　性能指标

为了衡量单核预测模型和本文混合核模型的性能优劣, 采用均方根误差 (e_{mse}), 平均绝对百分比误差 (e_{mape}) 和相关系数 (R^2) 3 个指标对模型进行评估:

$$e_{mse}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)},\quad e_{mape}=\frac{1}{n}\sum_{1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|,$$

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

式中: y_i 表示第 i 个样本对应的 $PM_{2.5}$ 浓度的真实值, $\hat{y}_i$ 表示第 i 个样本对应的 $PM_{2.5}$ 浓度的预测值, $\bar{y}$ 表示模型预测输出平均值. e_{mse} 反映模型预测输出值稳定性, e_{mape} 反映模型预测输出值偏离实际值的程度, 两者均是越小说明模型性能越好; R^2 反

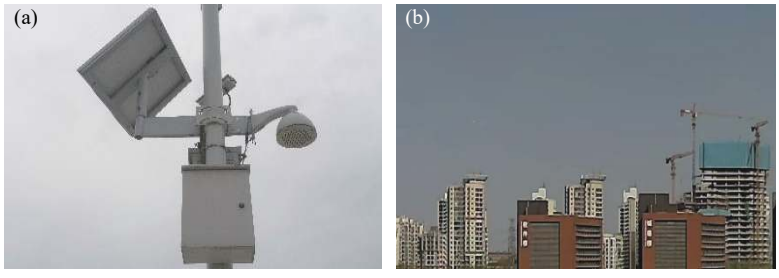

图 1　数据采集设备(a)及数据样本(b)

Fig.1　Data acquisition equipment (a) and data samples (b)

映模型预测输出值与真实值之间的关联程度，其值越接近 1 说明模型性能越好.

4 结果分析

本实验使用大气图像数据和对应的空气质量 $PM_{2.5}$ 数据进行实验. 图像数据来源于安装在北京工业大学内的 360 智能摄像头，采集 2019 年 1 月 1 日至 2019 年 5 月 31 日每日 9:00 ~ 16:00 的 600×320 图像（每小时采样）共 1000 幅. $PM_{2.5}$ 数据来自安装在北京工业大学校园内的 808 微型气象站. 数据采集设备及数据样本如图 1 所示.

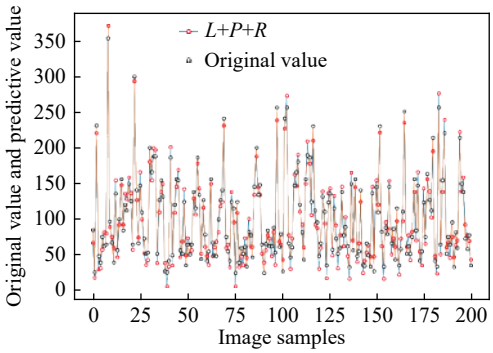

图 2 混合核模型预测值

Fig.2 Prediction results of mixture kernel model

4.1 预测结果分析

从采集的图像数据中随机抽取 600 张图像，将经过标准化处理的特征数据随机取 400 组作为训练样本，剩余 200 组作为测试样本. 为了证明基于图像混合核与列生成模型的有效性，将该模型与单核预测模型实验结果进行对比.

针对基于图像混合核的列生成 $PM_{2.5}$ 预测模型，利用预留的验证集数据进行仿真实验，仿真结果如图 2 和 3 所示. 从图 2 和 3 中可以看出，采用基于图像混合核的列生成模型对 1 h 后的 $PM_{2.5}$ 值进行预测，预测值与期望输出值基本相吻合，能达

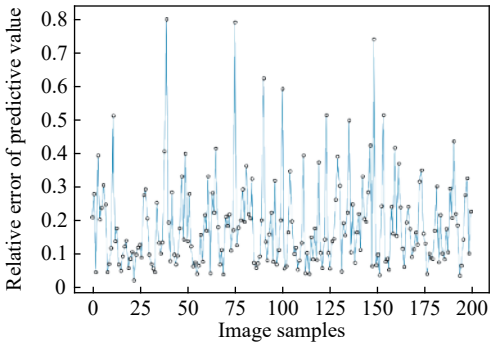

图 3 混合核模型预测相对误差

Fig.3 Relative error in mixture kernel model prediction

到相对较高的预测精度. 预测值的相对误差绝大部分维持在较低范围内.

除了个别因环境因素、人为因素等造成的较大偏差外，基本可以认为该模型满足了预测精度要求. 同时，将基于图像混合核的列生成 $PM_{2.5}$ 预测模型与单核预测模型进行对比实验，结果如图 4 所示.

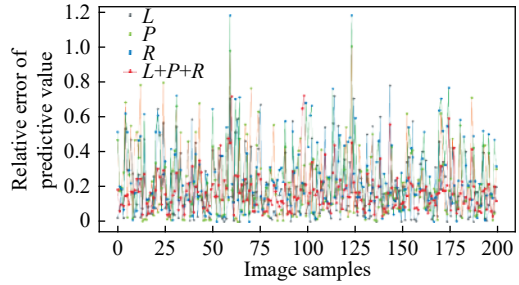

图 4 4 种模型预测相对误差

Fig.4 Relative error in prediction for the four models

从图 4 中可以看出，对于同一测试样本，基于图像混合核的列生成模型的预测相对误差要普遍小于其他单模型，且混合核模型的预测相对误差稳定维持在一定范围内，未出现较大误差，可认为基于图像混合核的列生成模型在预测性能和模型稳定性方面优于其他三个单核预测模型.

结合 3 个性能指标对 4 种预测模型进行对比，结果如表 2 所示. 相比于 3 种单核预测模型，基于图像混合核的列生成模型预测结果的均方根误差（e_{mse}）和平均绝对百分比误差（e_{mape}）最小，相关系数（R^2）最大，说明基于图像混合核的列生成模型表现出了更高的预测精度和预测稳定性.

表 2 4 种模型性能对比

Table 2 Performance comparison of the four models

Kernel	e_{mse}	e_{mape}/%	R^2
L	11.959	13.603	0.814
P	13.924	15.601	0.751
R	11.188	12.213	0.843
L+*P*+*R*	9.553	9.955	0.895

4.2 计算复杂度分析

基于图像混合核的列生成预测模型的计算复杂度取决于基于图像特征值的模型建立过程，因此其计算复杂度与列生成算法相等. 列生成算法的计算复杂度计算如下：设样本总数为 n，则混合核矩阵总列数为 np，最终要抽取 m 列. 抽取每列都要与其余所有列进行计算对比，则计算次数依次

为 $np, np-1, np-2, \cdots, np-m+1$, 总计算次数为 $mnp-\left(m^2-m\right)/2$. 因为 $m \ll n$, 所以混合核模型的计算复杂度可表示为 $O(nmp)$, 同理的单核预测模型计算复杂度为 $O(nk)$(k 为单核矩阵中抽取的列数). 本文中 $p=3, m \ll n, k \ll n$, 可得 $O(nmp)=O(nk)=O(n)$, 所以混合核预测模型与单核预测模型相比, 计算复杂度没有明显增加.

综上, 本文提出的基于图像混合核的列生成预测模型, 在满足预测精度的前提下, 获取数据的成本更低, 获取数据的途径更便捷, 计算复杂度与单核预测模型相比无明显增加, 对进行 $PM_{2.5}$ 预测有一定的借鉴意义.

5　结论

列生成算法是解决多变量线性规划问题的典型方法, 核函数可以将非线性数据映射到高维线性空间, 本文将核技巧与列生成算法相结合, 提出了一种基于图像混合核的列生成预测模型. 通过实验得出以下结论:

(1)针对大气 $PM_{2.5}$ 预测影响因素复杂、大气污染物浓度数据难以获取等问题, 基于图像数据建立模型进行预测是可行的, 当选取的图像特征与 $PM_{2.5}$ 密切相关时, 能够取得不错的预测效果.

(2)基于图像混合核的列生成预测模型无需考虑组合参数问题, 且能从核矩阵中选择最佳的列, 使模型的解具备稀疏性且预测精度可观.

(3)混合核模型比普通单核预测模型的预测误差小、精度高, 模型稳定性好, 该模型具备良好的预测性能.

(4)本文提出的模型对多雾、降雨和夜间等天气无法适用, 会影响模型预测效果, 需要在今后的工作中将此类特殊天气条件考虑到模型训练中, 期望得到泛化能力更强、预测精度更高的预测模型.

参　考　文　献

[1] Zhang X L, Zhao J H, Cai B. Prediction model with dynamic adjustment for single time series of $PM_{2.5}$. *Acta Automatica Sinica*, 2018, 44(10): 1790
(张熙来, 赵俭辉, 蔡波. 针对$PM_{2.5}$单时间序列数据的动态调整预测模型. 自动化学报, 2018, 44(10): 1790)

[2] Zhan Y, Luo Y Z, Deng X F, et al. Spatiotemporal prediction of continuous daily $PM_{2.5}$, concentrations across China using a spatially explicit machine learning algorithm. *Atmos Environ*, 2017, 155: 129

[3] Sun W, Sun J Y. Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *J Environ Manage*, 2016, 188: 144

[4] Qu Y, Qian X, Song H Q, et al. Machine-learning-based model and simulation analysis of $PM_{2.5}$ concentration prediction in Beijing. *Chin J Eng*, 2019, 41(3): 401
(曲悦, 钱旭, 宋洪庆, 等. 基于机器学习的北京市$PM_{2.5}$浓度预测模型及模拟分析. 工程科学学报, 2019, 41(3): 401)

[5] Russo A, Raischel F, Lind P G. Air quality prediction using optimal neural networks with stochastic variables. *Atmos Environ*, 2013, 79: 822

[6] 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 Beijing Univ Technol*, 2018, 44(12): 1494
(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. 北京工业大学学报, 2018, 44(12): 1494)

[7] Liu C B, Tsow F, Zou Y, et al. Particle pollution estimation based on image analysis. *PloS One*, 2016, 11(2): e0145955

[8] Gu K, Qiao J F, Li X L. Highly efficient picture-based prediction of $PM_{2.5}$ concentration. *IEEE Trans Ind Electron*, 2019, 66(4): 3176

[9] Li X L, Zhang S, Wang K. $PM_{2.5}$ air quality prediction based on image quality analysis. *J Beijing Univ Technol*, 2020, 46(2): 191
(李晓理, 张山, 王康. 基于图像质量分析的$PM_{2.5}$空气质量预测. 北京工业大学学报, 2020, 46(2): 191)

[10] Wang H Q, Sun F C, Cai Y N, et al. On multiple kernel learning methods. *Acta Autom Sin*, 2010, 36(8): 1037
(汪洪桥, 孙富春, 蔡艳宁, 等. 多核学习方法. 自动化学报, 2010, 36(8): 1037)

[11] Fink M, Desaulniers G, Frey M, et al. Column generation for vehicle routing problems with multiple synchronization constraints. *Eur J Oper Res*, 2019, 272(2): 699

[12] Li H. *Statistical Learning Method*. Beijing: Tsinghua University Press, 2012
(李航. 统计学习方法. 北京: 清华大学出版社, 2012)

[13] Demiriz A, Bennett K P, Shawe-Taylor J. Linear programming boosting via column generation. *Mach Learn*, 2002, 46(1-3): 225

[14] Bi J B, Zhang T, Bennett K P. Column-generation boosting methods for mixture of kernels//*Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, 2004: 521

[15] Vapnik V. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013

[16] Berman D, Treibitz T, Avidan S. Single image dehazing using haze-lines. *IEEE Trans Pattern Anal Mach Intell*, 2018, 42(3): 720

[17] Seinfeld J H, Pandis S N. Atmospheric Chemistry and Physics: from Air Pollution to Climate Change. John Wiley & Sons, 2016

[18] Graves N, Newsam S. Camera-based visibility estimation: Incorporating multiple regions and unlabeled observations. *Ecol*

Inform, 2014, 23: 62

[19] He K M, Sun J, Tang X O. Single image haze removal using dark channel prior. *IEEE Trans Pattern Anal Mach Intell*, 2011, 33(12): 2341

[20] Kim K W, Kim Y J. Perceived visibility measurement using the HSI color difference method. *J Korean Phys Soc*, 2005, 46(5): 1243

[21] Yuan L, Mu Z C, Liu L M. Ear recognition based on kernel principal component analysis and support vector machine. *J Univ Sci Technol Beijing*, 2006, 28(9): 890
(袁立, 穆志纯, 刘磊明. 基于核主元分析法和支持向量机的人耳识别. 北京科技大学学报, 2006, 28(9): 890)