DOI:10.3969/j.issn.1003-5060.2024.07.017

基于函数型 k 近邻分类模型的 $PM_{2.5}$ 研究

刘 壮, 凌能祥
(合肥工业大学 数学学院,安徽 合肥 230601)

摘 要:文章利用函数型数据分析方法,选取每天 24 h 的温度数据作为一条独立的曲线样本,并在该基础上建立函数型 k 近邻分类模型,用以对当天的 24 h 平均 $PM_{2.5}$ 质量浓度进行分类判别。分别选取二次型核函数、指数型核函数、三角型核函数建立 k 近邻分类模型,并对其结果进行分析,通过对比发现,利用三角型核函数的 k 近邻分类模型对 $PM_{2.5}$ 质量浓度进行分类的准确性最高且最稳健。采用 NW(Nadaraya-Watson)核方法与 k 近邻分类模型进行比较分析,结果表明,k 近邻分类模型能有效提高分类的准确率。

关键词:函数型数据分类;k 近邻;核函数;非参数统计
中图分类号:O212.7 **文献标志码**:A **文章编号**:1003-5060(2024)07-0967-04

Analysis of $PM_{2.5}$ based on functional k-nearest neighbors classification model

LIU Zhuang, LING Nengxiang
(School of Mathematics, Hefei University of Technology, Hefei 230601, China)

Abstract: In this paper, a functional data analysis method is used to select temperature data of 24 h per day as an independent curve sample. On this basis, a functional k-nearest neighbors(KNN) classification model is established to classify and discriminate average $PM_{2.5}$ concentration of the day. The quadratic kernel function, exponential kernel function, and triangle kernel function are selected to establish the kNN classification model, and the results are analyzed. Through comparison, it is found that the kNN classification model using triangle kernel function is the most accurate and robust in classifying $PM_{2.5}$ concentration. A comparative analysis is performed using the Nadaraya-Watson (NW) kernel method and the kNN classification model. The results show that the kNN classification model can effectively improve the classification accuracy.

Key words: functional data classification; k-nearest neighbors (KNN); kernel function; nonparametric statistics

0 引 言

随着我国经济的高速发展,居民生活水平已经得到显著提高,进而导致机动车数量迅速增加。另一方面,随着工业生产规模不断扩大,会消耗大量的化石燃料。因此,各种有害气体的排放导致我国空气质量明显下降。近年来,我国雾霾现象比较严重,空气中高质量浓度的 $PM_{2.5}$ 是形成雾霾天气的最主要原因之一,公众对于 $PM_{2.5}$ 等指标的关注度明显提高。$PM_{2.5}$ 是指环境中空气动力学当量直径小于等于 2.5 μm 的颗粒物。它能较长时间悬浮于空气中,其在空气中质量浓度越高,代表空气污染越严重。根据 2012 年我国发布的《环境空气质量标准》将 $PM_{2.5}$ 24 h 平均质量浓度限值分别定在 35、75 $\mu g/m^3$。气象因素和 $PM_{2.5}$ 质量浓度的变化具有较强的相关性,例如温度、湿度、风速等。而温度是影响 $PM_{2.5}$ 质量浓度主要的气象因素,因此研究温度与 $PM_{2.5}$ 质量浓度之间的关系具有重要意义。目前,许多学者研究了气象因素影响大气中 $PM_{2.5}$ 质量浓度问题。

收稿日期:2020-02-26;**修回日期**:2020-03-16
作者简介:刘 壮(1995—),男,安徽六安人,合肥工业大学硕士生;
凌能祥(1963—),男,安徽合肥人,博士,合肥工业大学教授,博士生导师.

文献[1]运用多重线性回归分析方法,研究了西安市莲湖区和雁塔区 $PM_{2.5}$ 质量浓度的变化特征及其与气象条件的关系,结果表明,两城区空气质量逐年改善,秋冬季 $PM_{2.5}$ 污染较为严重,气象因素影响大气中 $PM_{2.5}$ 质量浓度水平;文献[2]通过统计 2014-01—2016-12 上海市 $PM_{2.5}$ 质量浓度的日观测数据,对上海市 $PM_{2.5}$ 质量浓度变化特征及其对气候变化的响应进行了分析。

由于气象因素会随着时刻点的变化而变化,而每个时刻点的气象因素并不是独立存在的,它们之间存在一定的相关关系。同时采用传统的多元统计方法往往会带来维度灾难等问题,且会忽略数据的函数特性。基于此,文献[3]基于函数型非参数曲线判别方法,并提出一种新的非参数工具,用于研究曲线(被视为功能预测变量)与分类响应之间的关系;文献[4]提出基于核函数规则的函数型数据分类方法;文献[5]介绍了非参数函数型数据分析的理论和应用,其中包括非参数函数型数据的分类方法;文献[6]基于函数型数据视角对京津冀地区的污染特征进行研究;文献[7]利用函数型数据方法对中国空气质量进行预测和聚类;文献[8]提出不完全函数型协变量的分类方法,并将该方法运用于医学数据集上;文献[9]提出在独立场合下,随机缺失型函数型数据 k 近邻估计方法,并将其应用于基于温度对 $PM_{2.5}$ 的估计和预测。

本文从安徽省马鞍山市气象站台测得的 $PM_{2.5}$ 质量浓度和温度数据出发,首先对数据进行筛选,然后选取每天 24 h 的温度数据作为函数型数据的解释变量,$PM_{2.5}$ 质量浓度按照国家标准分为 2 类:$\rho(PM_{2.5})\leqslant 75\ \mu g/m^3$,即空气质量为优和良,相对空气质量较好的作为第 1 类;而 $\rho(PM_{2.5})>75\ \mu g/m^3$ 的作为第 2 类,即空气质量为差的情况。

最后建立函数型非参数 k 近邻分类模型,将样本数据导入模型并进行分类,对不同核函数的分类结果进行对比分析,选取准确性较高的分类模型,并与相同参数下的 NW(Nadaraya-Watson)核方法进行比较。

1 函数型 k 近邻分类模型

1.1 后验概率的 k 近邻估计量

设 $(\boldsymbol{\chi}_i,y_i)_{i=1,\cdots,n}$ 是总体 $(\boldsymbol{\chi},Y)$ 的 n 个独立同分布样本,取值于 $E\times\overline{G}=\{1,\cdots,G\}$,其中 (E,d) 是一个半度量向量空间(例如,$\boldsymbol{\chi}$ 是一个函数型随机变量,且 d 是一个半度量),则判别的后验概率 $p_g(\boldsymbol{\chi})=E(I_{[Y=g]}|\boldsymbol{\chi}=x)$,其中:

$$I_{[Y=g]}=\begin{cases}1, & Y=g;\\ 0, & Y\neq g。\end{cases}$$

后验概率的 k 近邻估计量可以按照条件期望表示,即

$$\hat{p}(x)=\hat{p}_{g,k}(x)=\frac{\sum\limits_{\{i;y_i=g\}}^{n}K\left(h_k^{-1}d(x,\boldsymbol{\chi}_i)\right)}{\sum\limits_{i=1}^{n}K\left(h_k^{-1}d(x,\boldsymbol{\chi}_i)\right)},$$

其中:K 为一个非对称核;h_k 为使得 $\mathrm{card}\{i;d(x,\boldsymbol{\chi}_i)<h_k\}=k$ 的窗宽。

为了达到最优分类准则,本文需要在有限集 $\{1,\cdots,k\}$ 中找到最优的 k 使得损失函数 $k_{\mathrm{Loss}}\leftarrow\underset{k\in\{1,\cdots,k\}}{\mathrm{argmin}}\ \mathrm{Loss}(k)$,$h_{\mathrm{Loss}}\leftarrow h_{k_{\mathrm{Loss}}}$ 达到最小,其中 Loss 损失函数是基于 $\hat{p}_{g,k}(\boldsymbol{\chi}_i)$ 和 y_i 的,本文可以选用错误分类率。错误分类率的计算公式为 $M=\frac{1}{n}\sum\limits_{i=1}^{n}I_{[y_i\neq y_i^{\mathrm{LCV}}]}$,其中 $I_{[y_i\neq y_i^{\mathrm{LCV}}]}$ 为示性函数(下同)。

1.2 k 近邻参数的选取

为了选取合适的参数 k,需要用交叉检验方法。主要目标是计算

$$p_g^{\mathrm{LCV}}(\boldsymbol{\chi})=\frac{\sum\limits_{\{i;y_i=g\}}^{n}K\left(d(\boldsymbol{\chi}_i,x)/h_{\mathrm{LCV}}(\boldsymbol{\chi}_{i_0})\right)}{\sum\limits_{i=1}^{n}K\left(d(\boldsymbol{\chi}_i,x)/h_{\mathrm{LCV}}(\boldsymbol{\chi}_{i_0})\right)},$$

其中:$i_0=\underset{i=1,\cdots,n}{\mathrm{argmin}}\ d(x,\boldsymbol{\chi}_i)$;$h_{\mathrm{LCV}}(\boldsymbol{\chi}_{i_0})$ 为通过交叉验证过程获得的对应于 $\boldsymbol{\chi}_{i_0}$ 最优近邻个数的窗宽。最优近邻个数 $k_{\mathrm{LCV}}(\boldsymbol{\chi}_{i_0})=\underset{k}{\mathrm{argmin}}\ \mathrm{LCV}(k,i_0)$,其中:

$$\mathrm{LCV}(k,i_0)=\sum_{g=1}^{G}[1_{[y_{i_0}=g]}-p_{g,k}^{(-i_0)}(\boldsymbol{\chi}_{i_0})]^2;$$

$$p_{g,k}^{(-i_0)}(\boldsymbol{\chi}_{i_0})=\frac{\sum\limits_{\{i;y_i=g,i\neq i_0\}}^{n}K\left(d(\boldsymbol{\chi}_i,\boldsymbol{\chi}_{i_0})/h_{k(\boldsymbol{\chi}_{i_0})}\right)}{\sum\limits_{\{i=1,i\neq i_0\}}^{n}K(d\left(\boldsymbol{\chi}_i,\boldsymbol{\chi}_{i_0})/h_{k(\boldsymbol{\chi}_{i_0})}\right)}。$$

1.3 核函数和半度量的选取

对于函数型 k 近邻分类模型,核函数和半度量的选取至关重要,可以供选择的非对称型核函数主要有:

1) 三角型核函数 $K(u)=(1-u)I_{[0,1]}(u)$;

2) 二次型核函数 $K(u)=\frac{3}{4}(1-u^2)I_{[0,1]}(u)$;

3) 示性核函数 $K(u)=I_{[0,1]}(u)$。

在本文研究过程中,将选取 3 种不同的核函

数进行比较分析。在实际应用中，半度量的种类主要有导数型半度量、主成分半度量和偏最小二乘半度量等，本文选取二阶导数型半度量 $d(\boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}) = \sqrt{\int [\hat{\boldsymbol{\chi}}_i^{(2)}(t) - \hat{\boldsymbol{\chi}}_i^{(2)}(t)]^2 \mathrm{d}t}$ 进行分析。

2　实证分析

2.1　数据预处理

本文的温度数据和 $PM_{2.5}$ 质量浓度数据来源于安徽省气象局 2015—2019 年的马鞍山站的 $PM_{2.5}$ 监测仪器。其中由于机器部分缺测或者故障，导致某些数据异常或者缺测。在预处理阶段予以剔除，最终得到准确可用的 1 270 条温度曲线和对应的 $PM_{2.5}$ 质量浓度，其中 200 条温度曲线如图 1 所示。

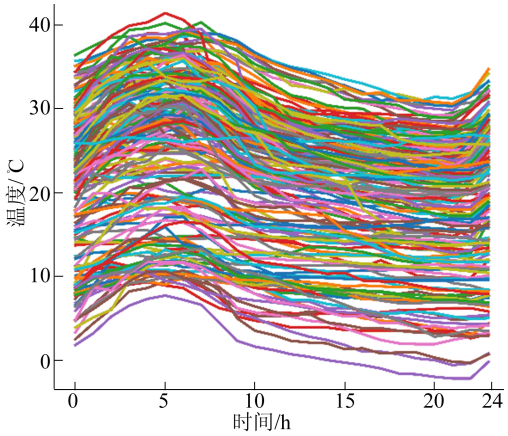

图 1　部分温度曲线

1 270 d 的日均 $PM_{2.5}$ 质量浓度折线图如图 2 所示。

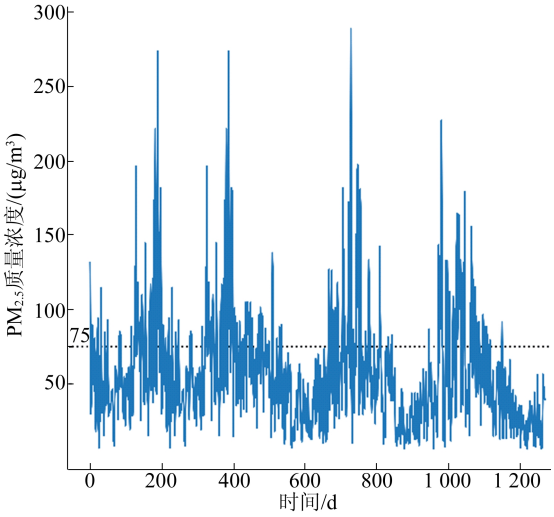

图 2　$PM_{2.5}$ 质量浓度折线图

2.2　函数型非参数二分类模型检验

2.2.1　k 近邻方法检验

将 $PM_{2.5}$ 质量浓度数据按照 75 $\mu g/m^3$ 的标准分为 2 类，然后作为标签对应原始的每日温度曲线数据，按照交叉验证的方法，随机生成 100 组 80%的训练集和 20%的测试集。本文基于二阶导数半度量，分别选取指示型核函数(Ind)、二次型核函数(Qua)和三角型核函数(Tri)，然后对 100 组样本数据集进行分类，并比较结果。对3 种不同的核函数对应导数半度量中 B 样条基不同节点个数进行 100 次分析判别的误分类比率箱线图如图 3 所示。

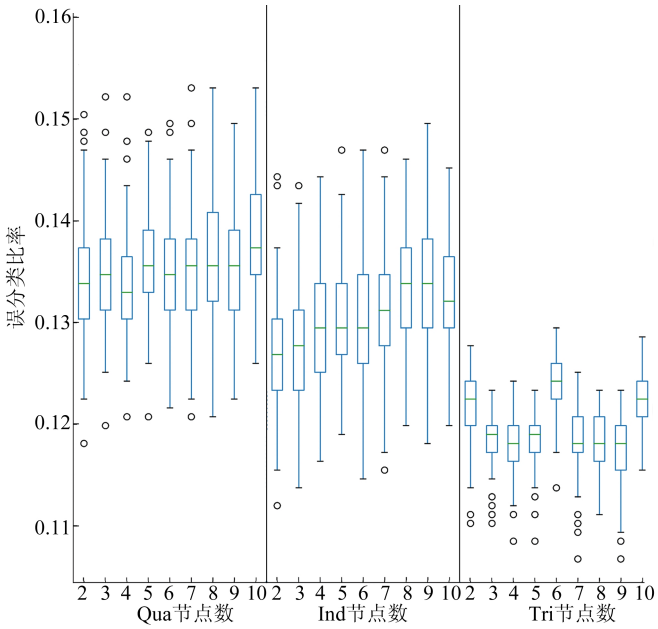

图 3　不同核函数和不同近邻数误分类率箱线图

从图3可以看出,在对$PM_{2.5}$进行判别分析的过程中,对交叉验证随机选取的数据集,3种核函数的表现都较为稳定。三角型核函数相对于二次型核函数和指示型核函数误分类比率相对更低,且对数据集的变化相对于另外2种方法更稳健。三角型核函数方法呈左偏态,而二次型核函数方法主要呈右偏态分布。3种核函数进行100次判别的误分类比率的均值见表1所列。

表1 不同核函数和不同近邻数100次训练平均误分类率

样条基节点数	2	3	4	5	6	7	8	9	10
二次型核函数	0.134 0	0.135 1	0.133 4	0.135 6	0.135 3	0.134 9	0.136 1	0.135 3	0.138 5
三角型核函数	0.127 1	0.127 6	0.129 1	0.130 3	0.130 2	0.131 2	0.133 6	0.134 1	0.132 8
指示型核函数	0.121 6	0.118 3	0.118 0	0.118 3	0.124 3	0.118 8	0.118 3	0.117 5	0.122 2

从表1可以看出:3种核函数的分类结果都相对较好,正确分类率都在86%以上,而三角型核函数的正确分类率在88%左右;二次型核函数和指示型核函数在B样条基节点数较小时表现得更为优良,整体误分类率趋势随着节点数的增加而增加,三角型核函数除了样条基节点数为2、6、10时误分类率略高以外,其他节点数对应的误分类率都较低且稳定。

2.2.2 NW核方法检验及2种方法结果比较

基于NW核方法,本文选用三角型核函数,同时将半度量也设置成二阶导数半度量,与k近邻分类模型保持一致。

将样本数据导入NW核方法模型,训练1次的结果见表2所列。

表2 NW核方法不同窗宽对应的误分类率

窗宽	7.84	8.14	8.44	8.75	9.05	9.35
误分类率	0.223 1	0.220 5	0.211 7	0.214 3	0.219 6	0.223 1
窗宽	9.65	9.95	10.26	10.56	10.86	11.16
误分类率	0.222 2	0.216 1	0.220 5	0.215 2	0.217 8	0.225 7

当采取交叉验证对模型训练100次时,NW核方法得到的平均误分类率为0.200 4。由表2可知,NW核方法对于本数据最优的误分类率大于k近邻分类模型的3种情况。k近邻分类模型相对于NW核方法,使用了局部窗宽,从而提高了分类的准确率。

3 结 论

函数型数据分析方法相对于传统多元方法在处理高维问题上具有优势,同时很好地避免了类似贝叶斯分类需要各个变量之间相互独立的条件。而非参数方法又避免了对于数据总体分布的假定,因此利用函数型非参数方法分析是非常合适的。与函数型NW核方法相比,k近邻分类模型考虑了局部性质,效果更好。本文的实际分析也证明了函数型数据非参数k近邻分类模型在$PM_{2.5}$分类上的可行性,为函数型数据的应用提供了新的方向。

[参 考 文 献]

[1] 孟昭伟,雷佩玉,张同军,等.2015—2018年西安市两城区PM_(2.5)质量浓度变化特征及气象影响因素[J].卫生研究,2020,49(1):75-79,85.

[2] 孙嘉文,方海羽.上海市PM_(2.5)浓度变化特征及其对气象变化的响应研究[J].安徽农学通报,2017,23(21):85-87.

[3] FERRATY F,VIEU P. Curves discrimination:a nonparametric functional approach[J]. Computational Statistics and Data Analysis,2003,44:161-173.

[4] ABRAHAM C,BIAU G,CADRE B. On the kernel rule for function classification[J]. Ann Inst Statist Math,2005,58(3):618-639.

[5] FERRATY F,VIEU P. Nonparametric functional data analysis:theory and practices[M]. Berlin:Springer,2006.

[6] 梁银双,刘黎明.京津冀地区PM2.5污染特征的研究:基于函数型数据分析的视角[J].运筹学学报,2018,22(2):105-114.

[7] 李海蓉.函数型数据视角下中国空气质量的预测及聚类研究[D].南京:南京信息工程大学,2018.

[8] MOJIRSHEIBANI M,SHAW C. Classification with incomplete functional covariates[J]. Statistics & Probability Letters,2018,139:40-46.

[9] 程彦茹.基于随机缺失函数型非参数/半参数模型的k近邻估计[D].合肥:合肥工业大学,2019.

(责任编辑 闫杏丽)