DOI: 10.11992/tis.201709029
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.tp.20180416.1311.006.html

一种加入类间因素的曲线聚类算法

许腾腾, 王瑞, 黄恒君
(兰州财经大学 统计学院, 甘肃 兰州 730020)

摘 要: 针对目前的曲线聚类算法基于类内差异设计, 造成不同类之间的曲线区分度不高的问题。在曲线拟合、曲线距离界定的基础上, 构造新的目标函数, 提出同时考虑类内和类间差异的曲线聚类算法。模拟结果显示, 该曲线聚类能够提高聚类精度; 针对 NO_2 污染物小时浓度的实例分析表明, 该曲线聚类算法具有更好的类间区分度。

关键词: 函数型数据; 类间差异; 曲线聚类; B-样条; 距离度量
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673−4785(2019)02−0362−07

中文引用格式: 许腾腾, 王瑞, 黄恒君. 一种加入类间因素的曲线聚类算法 [J]. 智能系统学报, 2019, 14(2): 362–368.
英文引用格式: XU Tengteng, WANG Rui, HUANG Hengjun. Curve clustering algorithms by adding the differences among clusters[J]. CAAI transactions on intelligent systems, 2019, 14(2): 362–368.

Curve clustering algorithms by adding the differences among clusters

XU Tengteng, WANG Rui, HUANG Hengjun
(School of Statistics, Lanzhou University of Finance and Economics, Lanzhou 730020, China)

Abstract: With the improvement of accuracy and frequency of data collection, functional data has appeared. Curves' clustering is a fundamental exploratory task in functional data analysis, and To sovave currently curves clustering algorithms available are based on the differences within each cluster, which has resulted in a low distinction among different curves. Therefore, on the base of curve fitting and curve distance, and with constructed objective function, curves clustering algorithms will be put forward with the consideration of cluster differences. Simulated results show that the curve cluster improves clustering accuracy. The example analysis of hourly NO_2 concentration ($\mu g/m^3$) indicates that this kind of curves clustering algorithms has a better distinction among different clusters.

Keywords: functional data; differences among clusters; curve clustering; B-spline; distance metric

随着信息技术的不断发展, 数据获取越来越便捷, 数据采集的密集化程度也越来越高。随之出现了一种具有函数特征的数据类型。如心理学研究中的脑电信号数据、生物技术中的基因微序列数据、化学计量中的光谱分析数据、经济研究中的股票分时成交价数据、环境监测中的污染物浓度数据等, 均随着时间变化而表现出明显的曲线特征。当前文献中将这种数据类型称为函数型数据 (functional data)[1]。

收稿日期: 2017−09−15. **网络出版日期**: 2018−04−18.
基金项目: 国家社科基金青年项目 (14CTJ009, 15CTJ004); 全国统计科学研究重点项目 (2017LZ43).
通信作者: 黄恒君. E-mail: noahwong@163.com.

一般而言, 函数型数据的曲线形式无法直接获取, 通常仅能够观测到其离散样本点, 并针对离散数据进行传统多元统计分析。当然, 这种做法由于未能考虑到数据的函数特性 (如连续、可导等), 同时需要处理高维问题, 往往不能取得很好的分析效果[2]。为此, 针对数据的曲线特征, 人们提出了各种分析方法[3-4], 包括函数型主成分分析、函数型线性模型、函数型聚类分析等, 将有限维的多元分析推广到无限维的函数型数据上来。

聚类分析是数据探索、数据压缩和展现的重要工具, 本文讨论函数型数据的聚类算法。目前, 函数型数据聚类分析方法大致分为两类[3]: 一是原始数据法, 该类方法直接针对离散样本点进

行聚类, 属于高维数据分析方法, 文献 [5] 对这种做法进行了综述。二是投影方法, 即以有限维的基底函数逼近曲线, 将无限维的问题转化为有限维问题展开分析。依据对基底函数所对应的系数处理方式不同, 又可分为滤波法和自适应法。滤波法将基底函数所对应的系数设定为固定参数, 分曲线拟合和聚类分析两步展开: 首先以有限维基底拟合曲线, 对估计的参数执行传统聚类算法。包括利用自组织映射 (SOM) 算法进行聚类和拟合函数型数据[6]; 利用两阶段随机过程分别完成数据降维和聚类[7] 等。根据基底函数选择利用 B-样条基底函数拟合数据并根据传统聚类方法分析[8-9], 利用正交基函数进行聚类分析[10] 等。自适应法是将基底函数所对应的系数作为随机变量处理, 将曲线拟合和聚类分析纳入一个目标函数, 采用类似 EM 的算法, 同时进行优化。如利用机器学习和神经网络模型 SVR 分析时空数据[11]、利用 STM 算法对时空数据进行聚类[12] 以及时间序列数据[13]、经维度数据[14] 等的聚类方法; 基于 K-medoids 项目聚类的协同过滤推荐算法[15]; 基于多元函数型主成分分析 (FPCA) 方法进行的改进混合模型同时进行曲线拟合与聚类分析[16]。在随机过程中利用 K-L 散度度量, 采用类似于 EM 算法进行聚类的算法[17] 等。

尽管有众多其他的算法[6, 18], 目前的函数型聚类分析仅考虑了类内部的差异, 而忽视了类间的差异性。对传统离散数据的聚类研究表明[19], 同时考虑类内与类间差异有助于提升聚类效果。

受此启发, 本文提出一种加入类间因素的曲线聚类算法。本文的做法属于滤波法, 包括 B-样条基底拟合曲线、曲线距离确定、曲线聚类目标函数设定, 以及加入类间因素的曲线聚类算法等。

1　加入类间因素的曲线聚类

根据前面的描述, 本文讨论的曲线聚类分析模型构建主要包含 3 个方面: 1) 由观测离散数据生成函数型数据, 这里采用 B-样条基底表述的方法; 2) 构造曲线之间的“距离”的表述, 并通过用 B-样条基底系数, 将曲线距离转化为传统欧氏距离; 3) 以所构造的距离作为亲疏程度度量, 并构建同时考虑类内差异和类间差异的目标函数, 完成曲线聚类任务。

1.1　曲线生成

假定数据 $\boldsymbol{Y}_i=[y_{i1}\ y_{i2}\ \cdots\ y_{im}]^{\mathrm{T}}(i=1,2,\cdots,n)$ 由如下模型生成:

$$y_{ij}=X_i(t_{ij})+\varepsilon_{ij},\quad j=1,2,\cdots,m \tag{1}$$

式中: $X_i(t)$表示待估计曲线, ε表示服从零均值同方差的独立同分布随机变量。进一步假定$X_i(t)$可由一组基底$\{\phi_{i1}(t),\phi_{i2}(t),\cdots\}$表示, 则有

$$X_i(t)=\sum_{k=1}^{\infty}\alpha_{ik}\phi_{ik}(t) \tag{2}$$

称这种做法为基底函数法, 它是一种将离散数据转化为曲线的常用平滑技术[3]。对待估计曲线$X_i(t)$采取截断处理, 得到如式 (3) 的形式:

$$X_i(t)=\sum_{k=1}^{K}\boldsymbol{\alpha}_{ik}\phi_{ik}(t)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Phi}_i(t) \tag{3}$$

从而将无限维问题转化为有限维估计方式。进一步假定[9]:

1) 对不同曲线$X_i(t)(i=1,2,\cdots,n)$采用一组相同的基底表述;

2) 基底函数设定为等距节点 B-样条基底。有

$$X_i(t)=\sum_{k=1}^{L}\alpha_{ik}B_{k,M}(t)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{B}_M(t) \tag{4}$$

式中: $L=K+M$, $B_{k,M}(t)$表示第k个内部节点数量为K的M阶 B-样条基底函数。$\boldsymbol{B}_M(t)$表示 M 阶 B-样条基底函数。对于参数$\boldsymbol{\alpha}_i$, 我们利用最小二乘法进行估计。

1.2　曲线距离

假定曲线$X_i(t)$为L^2空间的元素。则根据L^2范数定义, 有曲线$X_i(t)$和$X_j(t)$的距离为

$$d^2(i,j)=\left\|X_i-X_j\right\|^2 \tag{5}$$

其中$\|\cdot\|$表示 L^2 范数, 由假定 1) 及式 (4) 知

$$X_i(t)-X_j(t)=[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]^{\mathrm{T}}\boldsymbol{B}_M(t) \tag{6}$$

结合式 (6), 式 (5) 可转化为

$$d^2(i,j)=[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]^{\mathrm{T}}\boldsymbol{K}[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j] \tag{7}$$

式中

$$\boldsymbol{K}=\begin{bmatrix}\langle B_1,B_1\rangle & \langle B_1,B_2\rangle & \cdots & \langle B_1,B_L\rangle\\ \langle B_2,B_1\rangle & \langle B_2,B_2\rangle & \cdots & \langle B_2,B_L\rangle\\ \cdots & \cdots & \ddots & \cdots\\ \langle B_L,B_1\rangle & \langle B_L,B_2\rangle & \cdots & \langle B_L,B_L\rangle\end{bmatrix}$$

其中, $\boldsymbol{K}$ 为$L\times L$实对称矩阵, $\boldsymbol{K}$中的每一个元素$\langle B_i,B_j\rangle$表示L^2空间的内积。但是类似于$d^2(i,j)=[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]^{\mathrm{T}}[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]$这种形式的距离公式并不适用于非正交基底函数[9], 为将曲线距离用传统距离公式表示, 对$\boldsymbol{K}$作楚列斯基 (Cholesky) 分解得$\boldsymbol{K}=\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$, 其中$\boldsymbol{L}$为上三角矩阵, 并令$\boldsymbol{b}_i=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_i$, 式 (7) 可表示为

$$\begin{aligned}d^2(i,j)&=[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]^{\mathrm{T}}\boldsymbol{K}[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]=\\ &[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}[\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j]=\\ &[\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_i-\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_j]^{\mathrm{T}}[\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_i-\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_j]=[\boldsymbol{b}_i-\boldsymbol{b}_j]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_j]\end{aligned} \tag{8}$$

需要说明的是, 式 (8) 完成了从曲线距离到一般距离的转变, 构成了将曲线聚类转化为传统多

元聚类问题的基础。利用式 (8), 运用传统聚类算法对$\boldsymbol{b}_i$进行聚类, 得到P类, 记为$i \in G_p (p \in 1,2,\cdots,P)$。

由$\boldsymbol{b}_i=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_i$得到$\boldsymbol{B}=\boldsymbol{AL}$, 其中$\boldsymbol{A}=[\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \cdots\ \boldsymbol{\alpha}_{n_p}]^{\mathrm{T}}$, $\boldsymbol{B}=[\boldsymbol{b}_1\ \boldsymbol{b}_2\ \cdots\ \boldsymbol{b}_{n_p}]^{\mathrm{T}}$。$n_p$表示第$G_p$类中的曲线数量, 令$\bar{X}(t_0)$表示随机选取的一条曲线作为初始类中心, $\bar{X}^{(G_p)}(t)$表示第G_p类中的类中心。则有$\bar{X}^{(G_p)}(t)=n_p^{-1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{BL}^{-1}\boldsymbol{B}_M(t)$。

1.3　改进的曲线聚类算法

聚类分析的目的是将同类型数据进行归类, 同时对不同类型的数据进行区分。文献 [19] 针对传统离散数据提出的 K-means 聚类扩展方法兼顾了类内、类间差异。具体来讲, 通过对数据集引入全局中心点实现类内差异最小化的同时类中心与全局中心点距离最大化。相比于 K-means 算法, 这种做法提高了聚类效果[19]。

受此启发, 本文将 K-means 聚类分析扩展到函数型聚类分析上。本文的曲线聚类目标函数为

$$F(\boldsymbol{\Phi},\boldsymbol{U})=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{\left\|X_i(t)-\bar{X}(t)\right\|^2}{\left\|\bar{X}(t)-\bar{X}(t_0)\right\|^2} \tag{9}$$

式中: $\boldsymbol{\Phi}$表示待估参数矩阵 ($\boldsymbol{A}$ 或 $\boldsymbol{B}$), $\boldsymbol{U}$ 表示由 u_{ik} 构成的矩阵, 其中$u_{ik}\in\{0,1\}$, $\sum_{K=1}^{K}u_{ik}=1$, $X_i(t)$表示曲线, $\bar{X}(t_0)$表示随机选取的一条曲线作为初始类中心, 结合式 (4) 的曲线基底表述, 得到目标函数:

$$F(\boldsymbol{\Phi},\boldsymbol{U})=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{\left\|\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{B}_M(t)-\bar{X}(t)\right\|^2}{\left\|\bar{X}(t)-\bar{X}(t_0)\right\|^2} \tag{10}$$

根据前面关于曲线距离的描述将式 (7)~(8) 代入式 (10) 得到

$$F(\boldsymbol{\Phi},\boldsymbol{U})=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{[\boldsymbol{b}_i-\boldsymbol{b}_*]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_*]}{[\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]} \tag{11}$$

式中: $\boldsymbol{b}_*=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_*$, $\boldsymbol{b}_0=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_*$表示第 k 类类中心对应的参数, $\boldsymbol{\alpha}_0$表示初始类中心曲线的参数。

目标函数确定后, 式 (11) 中含有两个未知参数$\boldsymbol{\alpha}$及$\boldsymbol{U}$。通过固定一项求解另一项的步骤来求解式 (11), 即

1) 固定$\boldsymbol{\Phi}=\hat{\boldsymbol{\Phi}}$, 求解函数$F\left(\hat{\boldsymbol{\Phi}},\boldsymbol{U}\right)$;

2) 固定$\boldsymbol{U}=\hat{\boldsymbol{U}}$, 求解函数$F\left(\boldsymbol{\Phi},\hat{\boldsymbol{U}}\right)$。

针对 1), 为使目标函数式 (11) 达到最小, 当目标函数分子中曲线与对应类中心曲线距离小时u_{ik}=1, 否则为 0, 即

$$u_{ik}=\begin{cases}1, & \sum_{i=1}^{n}\sum_{k=1}^{K}\dfrac{[\boldsymbol{b}_i-\boldsymbol{b}_*]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_*]}{[\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]}\leqslant\boldsymbol{D}\\ 0, & \text{其他}\end{cases} \tag{12}$$

式中: $\boldsymbol{D}=\sum_{i=1}^{n}\sum_{k=1}^{K}\dfrac{[\boldsymbol{b}_i-(\boldsymbol{b}_*)']^{\mathrm{T}}[\boldsymbol{b}_i-(\boldsymbol{b}_*)']}{[(\boldsymbol{b}_*)'-\boldsymbol{b}_0]^{\mathrm{T}}[(\boldsymbol{b}_*)'-\boldsymbol{b}_0]}$, 且$(\boldsymbol{b}_*)'\neq(\boldsymbol{b}_*)$。

针对 2), 假设$\boldsymbol{b}_*$已知, 对目标函数式 (11) 关于$\boldsymbol{b}_*$求偏导数:

当$\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\left([\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]\right)^2=0$ 时, 令$\boldsymbol{b}_*=\boldsymbol{b}_0$,

当$\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\left([\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]\right)^2\neq0$时, 目标函数式 (11) 关于$\boldsymbol{b}_*$, 求导

$$\frac{\partial F\left(\boldsymbol{\Phi},\hat{\boldsymbol{U}}\right)}{\partial\boldsymbol{b}_*}=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{\partial\left(\dfrac{[\boldsymbol{b}_i-\boldsymbol{b}_*]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_*]}{[\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]}\right)}{\partial\boldsymbol{b}_*}=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{2[\boldsymbol{b}_i-\boldsymbol{b}_*]}{[\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]}-\frac{2[\boldsymbol{b}_*-\boldsymbol{b}_0][\boldsymbol{b}_i-\boldsymbol{b}_*]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_*]}{\left([\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]\right)^2}=0$$

得出

$$\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{2[\boldsymbol{b}_i-\boldsymbol{b}_*]}{[\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]}=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\frac{2[\boldsymbol{b}_*-\boldsymbol{b}_0][\boldsymbol{b}_i-\boldsymbol{b}_*]^{\mathrm{T}}[\boldsymbol{b}_i-\boldsymbol{b}_*]}{\left([\boldsymbol{b}_*-\boldsymbol{b}_0]^{\mathrm{T}}[\boldsymbol{b}_*-\boldsymbol{b}_0]\right)^2}$$

进一步化简得到$\boldsymbol{b}_*$

$$\boldsymbol{b}_*=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}[\boldsymbol{b}_i-\boldsymbol{b}_0]^{\mathrm{T}}\boldsymbol{b}_i}{\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}[\boldsymbol{b}_i-\mathbf{b}_0]^{\mathrm{T}}}$$

即

$$\boldsymbol{b}_*=\begin{cases}\boldsymbol{b}_0, & \sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\left[\left|\boldsymbol{b}_*\boldsymbol{b}_*^{\mathrm{T}}\right|\left(\boldsymbol{b}_*\boldsymbol{b}_*^{\mathrm{T}}\right)^{-1}\boldsymbol{b}_*-\boldsymbol{b}_0\right]=0\\ \dfrac{\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}[\boldsymbol{b}_i-\boldsymbol{b}_0]^{\mathrm{T}}\boldsymbol{b}_i}{\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}[\boldsymbol{b}_i-\boldsymbol{b}_0]^{\mathrm{T}}}, & \text{其他}\end{cases} \tag{13}$$

在进行计算机编程时可以不断对步骤 1)、2) 进行迭代, 直至找出最优$\boldsymbol{U}$和$\boldsymbol{\Phi}$。算法流程如下:

Input: $X=\{X_1,X_2,\cdots X_n\},k$

Initialize: Randomly choose an initial $\boldsymbol{b}_0=\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_k$

Repeat

Fixed $\boldsymbol{\Phi}$, use eq. (12) to solve$\boldsymbol{U}$

Fixed $\boldsymbol{U}$, use eq. (13) to solve$\boldsymbol{\Phi}$

Until convergence.

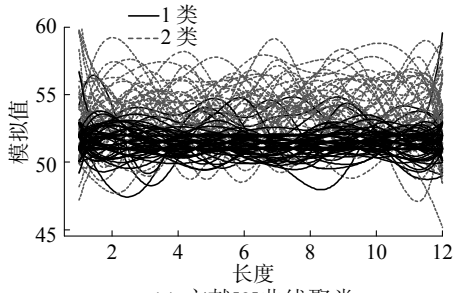
进一步, 由$\boldsymbol{b}_*=\boldsymbol{L}^{\mathrm{T}}\alpha_*$, 求解出$\boldsymbol{b}_*$可得到参数$\alpha_*$, 并根据式 (4) 还原出类中心曲线。

2　算法效果模拟验证与分析

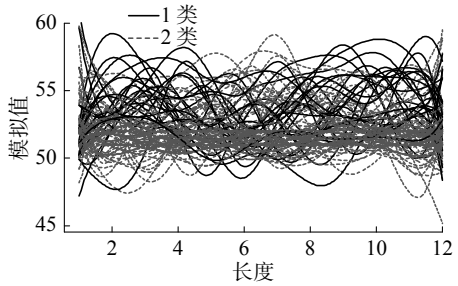
为验证本文曲线聚类算法的效果, 利用模拟数据与文献 [9] 中曲线聚类方法进行比较。模拟

数据由两组高斯分布生成两类曲线构成。模拟过程中两类高斯分布均值取 0.5 和 1, 方差取 0.7 和 1。在确定类别的前提下比较本文算法与文献 [9] 曲线聚类算法的聚类效果。聚类效果评价指标采用兰德指数 (Rand index) 评价算法的性能[20]。同时分析两组高斯分布的参数 (均值和方差) 对聚类的影响。分析结果显示: 同均值异方差情况下两种曲线聚类方法聚类结果均存在一定的误判, 异均值异方差情况下二者聚类也存在误判, 异均值同方差情况下二者聚类未出现误判。以下针对这一现象做出分析。

该部分采用 R 软件进行数据模拟分析, 每组包含n条数据, 每条数据含有m个数据点, 则模拟数据中每组高斯分布要生成$m\times n$个随机数。为保证拟合结果的光滑, 内部节点采用等距节点设置方式。针对高斯分布中的均值和方差分别在同均值异方差、同方差异均值、异均值异方差情况下分析本文的曲线聚类方法与已有曲线聚类方法的效果, 并对相应结果进行分析。为便于表述, 两类模拟数据分别记为 1 类和 2 类, 生成的区间长度设置为 12。为便于展示, 本文以图 1 异均值异方差条件下两种聚类方法比较为例。



(a) 文献[9]曲线聚类



(b) 本文曲线聚类

图 1　模拟数据曲线聚类对比

Fig. 1　Comparison with simulated data of curve's clustering

图 1 表明: 两组高斯分布参数不同条件下, 本文方法与文献 [9] 相比, 图 1(b) 中 1 类曲线分布密集程度大于图 1(a) 中 1 类曲线。为避免模拟次数少或其他原因对聚类效果的影响, 对 3 种类型的数据分别模拟一万次, 比较两种方法的平均错判率, 定义错判率=abs(1 类个数−n)/n, 模拟验证中m=12, n=50, 错判率下降比例=文献 [9] 方法错判率-本文方法错判率。结果见表 1。

表 1、2 表明: 无论本文的曲线聚类还是文献 [9] 中的曲线聚类方法, 类中心的变化与高斯分布中均值有关, 而聚类效果好坏与高斯分布的方差有关。对比表 1、2 中的同均值异方差和异均值异方差的错判率及兰德指数可以得出: 当两类高斯分布均值相同, 方差不同时, 两种方法对应的兰德指数相比于其他类型数据偏低。同时方差因素对聚类效果也会产生影响。综合比较表 1、2 中的 3 类数据错判率及兰德指数, 可以得到: 对于曲线聚类分析, 聚类效果会同时受数据总体均值和方差的影响, 对比分析表 1、2 均值相同方差不同的情形, 可以得到: 均值对聚类的影响程度要大于方差, 同时表 1、2 对两种方法错判率对比结果显本文的方法能够降低聚类错判率从而提高聚类效果。

表 1　3 种类型模拟数据平均错判率

Table 1　Average error rate of three types' simulated data

参数	文献[9]方法	本文方法	错判率下降比例
同均值异方差平均错判率 (mean=0.5,Var=1; mean=0.5,Var=0.7)	0.308	0.183	0.125
异均值同方差平均错判率 (mean=0.5,Var=1; mean=1,Var=1)	0.000	0.000	0.000
异均值异方差平均错判率 (mean=0.5,Var=1; mean=1,Var=0.7)	0.099	0.084	0.015

注: 错判率=abs(1类个数−n)/n, 模拟验证中m=12, n=50; 错判率下降比例为文献[9]方法错判率−本文方法错判率

表 2　3 种类型模拟数据兰德指数

Table 2　Rand index of three types' simulated data

参数	文献[9]方法	本文方法
同均值异方差兰德指数 (mean=0.5,Var=1;mean=0.5,Var=0.7)	0.740	0.780
异均值同方差兰德指数 (mean=0.5,Var=1;mean=1,Var=1)	1.000	1.000
异均值异方差兰德指数 (mean=0.5,Var=1;mean=1,Var=0.7)	0.850	0.870

3　NO_2 小时浓度曲线聚类效果分析

空气质量, 不仅关乎人类生存质量, 同时也是衡量可持续发展能力和宜居程度的重要指标。NO_2 是一种重要的机动车尾气污染物, 其污染程度涉及人们生活出行的健康。近年来, 空气质量

问题引起人们广泛的关注，大气污染监测数据成为人们了解空气质量的客观途径，也构成空气质量统计分析的数据基础。

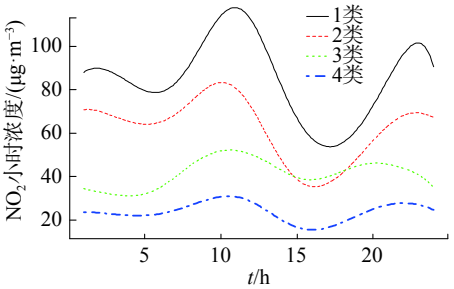
作为示例，通过实时网络爬虫手段[21]，采集兰州市铁路设计院空气质量监测站 (交通污染控制点) 的 NO_2 小时浓度数据，采用本文的曲线聚类算法展开大气污染等级聚类分析，并与传统曲线聚类结果进行比较。我们分析的样本期为 2013 年 6 月 1 日—10 月 14 日。

根据前面的方法，采用 B-样条基底函数进行曲线聚类分析。为保证拟合结果光滑，两种聚类方法样条基底阶数 M 均设置为 5，节点采用等距节点设置为 11(文中采用广义交叉验证准则进行节点数量选择)。考虑相同类中心下，与文献 [9] 曲线聚类进行聚类效果对比，如图 2 所示。

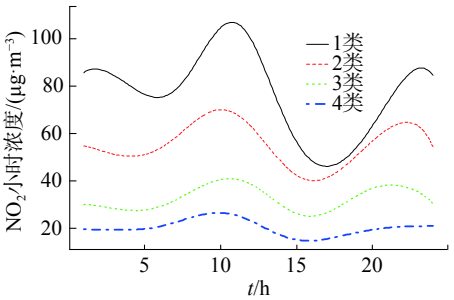
图 2 表明，K=5 时类中心聚类效果优于 K=4，即随着类中心个数的增加，两种方法的聚类效果均有所提升，说明类中心个数的确定在曲线聚类中起到关键作用。但需要指出的是，本文方法的类中心分布曲线更为平滑，类间的类中心曲线分布更为分散，进一步说明本文提出的方法聚类效果优于已有聚类方法。此外，考虑到实际应用，可将图 2 中的不同类别曲线看作空气质量污染物等级划分[20]。对比图 2(a)、(c) 与图 2(b)、(d) 可以发现，在空气质量实时监测过程中，图 2(a)、(c) 出现不同等级交叉情况，这对空气质量等级划分及应对会造成影响[22]。图 2(b)、(d) 在进行空气质量分析过程中能够较好的对空气质量进行聚类。另外，相比于针对离散数据的传统 K-means 聚类分析[23]，本文方法能够实时检测 NO_2 小时浓度变化趋势，并依据该变化趋势对污染物进行等级划分。

为便于展示，本文以 K=5 的曲线聚类结果为例，结果见图 3。图 3 表明，相比于已有曲线聚类算法，利用本文曲线聚类算法类内曲线分布集中，类间差异化明显。这与图 2 中两种曲线聚类算法类中心比较结果相一致。说明本文方法具有较好的类间区分度。

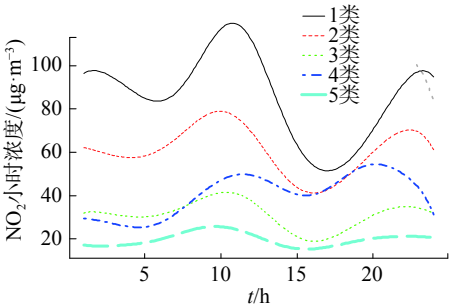
为进一步验证本文曲线聚类的聚类效果，对两种方法的分类精确度采用公式：类间差异/(类内差异+类间差异) 进行对比，见图 4。图 4 表明，随着类中心个数的增加，两种曲线聚类算法聚类效果均有所提高。本文曲线聚类的聚类效果要好于文献 [9] 的方法。通过与文献 [9] 方法进行比较，本文方法在 4 类的聚类效果低于 3 类聚类效果，随着类中心个数大于 4 类，聚类效果才逐步随着类中心个数增加聚类效果不断提升。说明本文



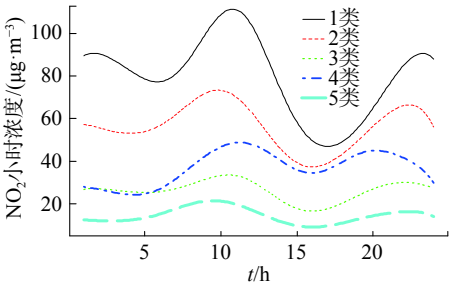
(a) 文献[9]曲线聚类, K=4



(b) 本文曲线聚类, K=4



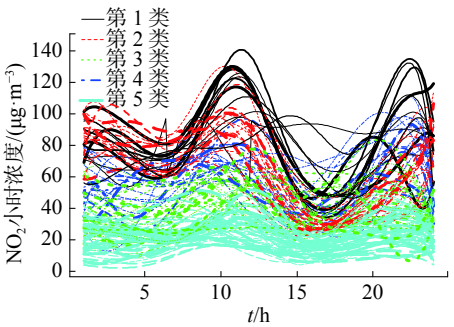
(c) 文献[9]曲线聚类, K=5



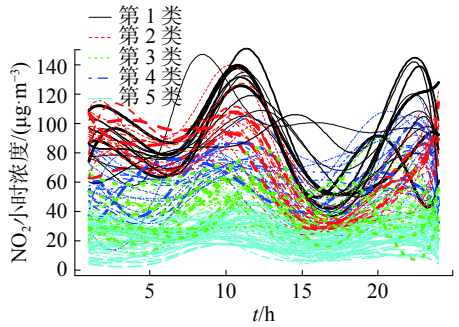
(d) 本文曲线聚类, K=5

图 2　曲线聚类类中心对比

Fig. 2　Comparison with curve cluster's center generated by different algorithms



(a) 文献[9]曲线聚类算法



(b) 本文算法

图 3　NO_2 小时浓度数据曲线聚类对比

Fig. 3　Comparison with curve clustering of NO_2 concentration

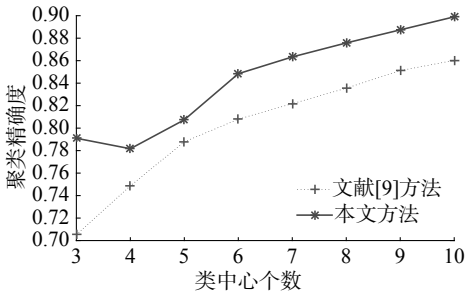


图 4　聚类效果对比结果

Fig. 4　Comparison with clustering effects

方法存在一定的不稳定性。

4　结束语

本文基于已有曲线聚类方法, 针对聚类效果不明显的问题, 提出加入类间因素的扩展曲线聚类算法。加入类间因素能够同时保证两类数据类内差异较小和类间差异较大。模拟数据及实例应用表明, 本文的曲线聚类算法有助于提高聚类效果。

需要说明的是, 本文的目的是将同时考虑类内和类间差异的做法引入曲线聚类算法。但我们的做法属于两步法, 即首先拟合曲线, 然后进行聚类。这种做法很难达到两部分的统一优化[24]。为此, 后续的工作是, 在同时考虑类内和类间差异的情况下, 进行自适应算法研究, 即将曲线拟合和聚类分析纳入一个目标函数, 同时进行优化。

参考文献:

[1] RAMSAY J O. When the data are functions[J]. Psychometrika, 1982, 47(4): 379–396.

[2] JACQUES J, PREDA C. Functional data clustering: a survey[J]. Advances in data analysis and classification, 2014, 8(3): 231–255.

[3] RAMSAY J O, SILVERMAN B W. Functional data analysis[M]. 2nd ed. New York: Springer, 2005: 1–18.

[4] FERRATY F, VIEU P. Nonparametric functional data analysis: theory and practice[M]. New York: Springer, 2006: 11–18.

[5] BOUVEYRON C, BRUNET-SAUMARD C. Model-based clustering of high-dimensional data: a review[J]. Computational statistics & data analysis, 2014, 71: 52–78.

[6] ROSSI F, CONAN-GUEZ B, GOLLI A E. Clustering functional data with the SOM algorithm[C]//proceedings of European Symposium on Artificial Neural Networks. Bruges, Belgium, 2004: 305–312.

[7] PENG Jie, MüLLER H G. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions[J]. The annals of applied statistics, 2008, 2(3): 1056–1077.

[8] ABRAHAM C, CORNILLON P A, MATZNER-LØBER E, et al. Unsupervised curve clustering using B-splines[J]. Scandinavian journal of statistics, 2003, 30(3): 581–595.

[9] 黄恒君. 基于 B-样条基底展开的曲线聚类方法 [J]. 统计与信息论坛, 2013, 28(9): 3–8.
HUANG Hengjun. Curves clustering using B-splines expansion[J]. Statistics & information forum, 2013, 28(9): 3–8.

[10] KAYANO M, DOZONO K, KONISHI S. Functional cluster analysis via orthonormalized gaussian basis expansions and its application[J]. Journal of classification, 2010, 27(2): 211–230.

[11] 王永坤, 王海洋, 潘平峻, 等. 面向公共安全的时空数据挖掘综述 [J]. 重庆邮电大学学报 (自然科学版), 2018, 30(1): 40–52.
WANG Yongkun, WANG Haiyang, PAN Pingjun, et al. A survey of data mining on spatial-temporal user behavior data for public safety[J]. Journal of chongqing university of posts and telecommunications (natural science edition), 2018, 30(1): 40–52.

[12] CHEAM A S M, MARBAC M, MCNICHOLAS P D. Model-based clustering for spatiotemporal data on air quality monitoring[J]. Environmetrics, 2017, 28(3): e2437.

[13] BOUVEYRON C, JACQUES J. Model-based clustering of time series in group-specific functional subspaces[J]. Advances in data analysis and classification, 2011, 5(4): 281–300.

[14] CHIOU J M, LI Pailing. Functional clustering and identifying substructures of longitudinal data[J]. Journal of the royal statistical society series B, 2007, 69(4): 679–699.

[15] 王永, 万潇逸, 陶娅芝, 等. 基于 K-medoids 项目聚类的协同过滤推荐算法 [J]. 重庆邮电大学学报 (自然科学版), 2017, 29(4): 521–526.
WANG Yong, WAN Xiaoyi, TAO Yazhi, et al. Collaborative filtering recommendation algorithm based on K-

medoids item clustering[J]. Journal of Chongqing university of posts and telecommunications (natural science edition), 2017, 29(4): 521–526.

[16] JACQUES J, PREDA C. Model-based clustering for multivariate functional data[J]. Computational statistics & data analysis, 2014, 71: 92–106.

[17] JACQUES J, PREDA C. Funclust: a curves clustering method using functional random variables density approximation[J]. Neurocomputing, 2013, 112: 164–171.

[18] 卞则康, 王士同. 基于混合距离学习的鲁棒的模糊 C 均值聚类算法 [J]. 智能系统学报, 2017, 12(4): 450–458.
BIAN Zekang, WANG Shitong. Robust FCM clustering algorithm based on hybrid-distance learning[J]. CAAI transactions on intelligent systems, 2017, 12(4): 450–458.

[19] HUANG Xiaohui, YE Yunming, ZHANG Haijun. Extensions of kmeans-type algorithms: a new clustering framework by integrating intracluster compactness and intercluster separation[J]. IEEE transactions on neural networks and learning systems, 2014, 25(8): 1433–1446.

[20] JAIN A K, DUBES R C. Algorithms for clustering data[M]. Upper Saddle River, NJ: Prentice-Hall, 1988: 227–229.

[21] 黄恒君, 漆威. 海量半结构化数据采集、存储及分析--基于实时空气质量数据处理的实践 [J]. 统计研究, 2014, 31(5): 10–16.
HUANG Hengjun, QI Wei. Massive semi-structured data: collection, storage and analysis--based on the practice of real-time air quality data processing[J]. Statistical research, 2014, 31(5): 10–16.

[22] 刘杰, 杨鹏, 吕文生, 等. 基于北京市 6 类污染物的环境空气质量评价方法 [J]. 安全与环境学报, 2015, 15(1): 310–315.
LIU Jie, YANG Peng, Lü Wensheng, et al. Environmental air quality evaluation method based on the six pollutants in the urban areas of Beijing[J]. Journal of safety and environment, 2015, 15(1): 310–315.

[23] 郭云飞, 林红飞, 郑旭. 中国城市空气质量指标的聚类分析 [J]. 统计与管理, 2016(8): 80–81.
GUO Yunfei, LIN Hongfei, ZHENG Xu. Clustering analysis of urban air quality indexes in China[J]. Statistics and management, 2016(8): 80–81.

[24] YAMAMOTO M, HWANG H. Dimension-reduced clustering of functional data via subspace separation[J]. Journal of classification, 2017, 34(2): 294–326.

作者简介：

许腾腾，男，1992 年生，硕士研究生，主要研究方向为异源异构数据整合与函数型数据分析。

王瑞，女，1993 年生，硕士研究生，主要研究方向为经济统计。

黄恒君，男，1981 年生，教授，博士，主要研究方向为异源异构数据整合与函数型数据分析。主持国家社会科学基金项目 1 项，获得省部级科研奖励 4 项。发表学术论文 30 余篇。