doi:10.3969/j.issn.1001-4616.2022.04.001

基于方向性多重假设检验和信息熵的函数型数据聚类新方法

杜秀丽,姜晓虎,孙晨曈,于　正

(南京师范大学数学科学学院,江苏 南京 210023)

[**摘要**]　近年来,针对函数型数据的聚类分析得到了一定程度的发展. 但当数据属于无限维空间时,会给聚类带来一定的难度. 传统聚类方法的局限性在函数型数据的聚类过程中日益凸显. 因此,本文提出了一种针对函数型数据的新聚类方法,能够更好地适应数据的特点,实现较好的聚类效果. 首先基于错误发现率的方向性多重假设检验和信息熵的理论,提出了新的平行度统计量,用以描述函数型曲线的形态差异. 在此基础上提出了新接近度的计算公式,最终改进了凝聚式层次聚类算法. 新的聚类方法被应用到 4 个不同类型的函数型数据集中,并与现有的其它方法的聚类结果进行分析和比较,证明了改进后的凝聚式层次聚类方法的有效性.

[**关键词**]　函数型数据聚类分析,错误发现率,方向性多重假设检验,信息熵,平行度
[**中图分类号**]O212.1　[**文献标志码**]A　[**文章编号**]1001-4616(2022)04-0001-09

A New Functional Data Clustering Method Based on Directional Multiple Hypothesis Test and Information Entropy

Du Xiuli, Jiang Xiaohu, Sun Chentong, Yu Zheng

(School of Mathematical Sciences, Nanjing Normal University, Nanjing 210023, China)

Abstract: In recent years, clustering analysis for functional data has been developed to a certain extent. However, when the data belong to infinite dimensional space, it will bring some difficulty to clustering. The limitations of traditional clustering methods are increasingly prominent in the clustering process of functional data. Therefore, this paper proposes a new clustering method for functional data, which can better adapt to the characteristics of data and achieve better clustering effect. Firstly, based on the directional multiple hypothesis test of false discover rate and the theoretical basis of information entropy, a new parallelism statistic is proposed to describe the morphological differences of functional curves. On this basis, a new calculation formula of proximity is proposed, and finally the condensed hierarchical clustering algorithm is improved. The new clustering method is applied to four different types of functional data sets, and the clustering results are analyzed and compared with other existing methods, which proves the effectiveness and advantages of the improved condensed hierarchical clustering method.

Key words: functional data clustering analysis, false discovery rate, directional multiple hypothesis testing, the information entropy, parallelism

近些年,函数型数据分析得到快速发展,而对其进行聚类分析,能够帮助我们更好地发现其中的潜在规律.到目前为止,统计学家提出许多函数型数据的聚类方法. 大致分为 4 类. 最简单的一类是原始数据聚类法,即在时间点上直接使用函数型数据的离散化形式. 此时,不必构建数据的函数形式. 但由于离散化数据的规模通常比较大,所以需要使用高维数据的聚类分析技术,比如 k-mean 聚类算法和层次聚类等[1-4]. 第二种为两阶段聚类方法. 第一阶段称为降维,就是用有限维基函数(比如 B 样条)[5]或函数型主成分分析[6-7]等逼近函数型曲线;第二阶段是对降维之后的多元变量应用经典的聚类分析方法实现聚

收稿日期:2022-01-21.
基金项目:国家社会科学基金项目(21BTJ044).
通讯作者:于正,副研究员,研究方向:教育管理. E-mail:33022@ njnu.edu.cn

类. 第三种为非参数方法,也是大家熟悉的聚类分析方法. 通常先计算数据点之间的欧式距离或其他方式定义的距离,再使用常规的非参数聚类方法[2,8]. 第四种是基于模型的聚类方法. 这类方法假设数据来自于潜在概率分布的混合. Jacques 和 Preda 于 2013 年提出了基于主成分的函数型数据模型聚类方法,称之为 funclust 方法[7];另一种基于模型的函数型聚类方法,是使用基展开系数进行的,James 和 Sugar 称之为 fclust 方法[5].

此外,在聚类分析中常见的一个问题是聚类数量的选择. 主要是使用常见的贝叶斯信息准则(BIC),赤池信息量准则(AIC)等[6,9-10].

函数型非参数聚类方法通常是基于距离、导数距离或者它们的综合进行聚类分析的,因为避免了估计参数的过程,所以便于实现,但仍存在着一些局限. Zambom 和 Collazos[4] 在文章中讨论了基于距离聚类的不足之处,并提出了一种基于距离和平行度的新 k-means 方法,但该方法的聚类效果过于依赖初始类中心的选择. 本文受此启发,综合了原始数据聚类方法和基于模型的聚类方法,提出一种函数型数据聚类的新方法. 该方法基于函数型数据样本间的差异提出了模型,并假定函数型差异曲线服从高斯过程,通过多重网格算法搜寻最优参数解,最终得到平行性差异的统计量. 另外,直接计算样本间的欧式距离. 最终结合平行度和距离差异的两类信息,通过改进的凝聚式层次聚类算法,对函数型数据进行聚类分析.

1 基于方向性多重假设检验和信息熵的函数型数据聚类分析

对于函数型数据来说,往往需要比较两条曲线在不同时间点上的差异,或者在一副图像中识别明显不同于其它位置的区域. 但在实际应用中除了考虑差异大小,我们也要关注方向差异带来的影响以及对方向判别误差率的控制. 因此,为了确定函数型曲线显著差异的具体区域,Xu 等[11] 受到多重假设问题的启发,提出检测两个函数型曲线均值函数的方向性多重假设检验,由此引出针对不同错误类型的 FDR.

1.1 基于错误发现率的方向性多重假设检验[11]

1.1.1 方向性多重假设检验

设 $\{Y_q(t):q=1,2;t\in T\}$ 为两个函数型数据的曲线,考虑以下模型:

$$\begin{cases}Y_1(t)=\mu(t)+\mu_d(t)+\varepsilon_1(t),\\ Y_2(t)=\mu(t)+\varepsilon_2(t),\end{cases} \tag{1}$$

式中,$\mu(t)$ 和 $\mu_d(t)$ 是未知函数,$\varepsilon_1(t)$ 和 $\varepsilon_2(t)$ 是独立随机变量.

对于如上函数型数据,二者之间的差异由 $\mu_d(t)$ 体现. 我们不仅需要知道两条曲线间是否存在差异,而且还需了解具体的方向差异,即 $\mu_d(t)>0$ 或 $\mu_d(t)<0$,这就需要进行方向性多重假设检验.

对每一时间点 t,具体假设如下:

$$\begin{cases}H_0(t):|\mu_d(t)|\leqslant\Delta,\\ H_1(t):\mu_d(t)<-\Delta,\\ H_2(t):\mu_d(t)>\Delta,\end{cases} \tag{2}$$

式中,Δ 为给定常数,体现两个样本曲线感兴趣的差异大小. 用 $z(t)$ 表示在时间点 t 假设的真实状态. 当 $H_0(t)$ 成立时,$z(t)=0$;当备择假设 1 或者 2 成立时,$z(t)=1$ 和 $z(t)=2$. 设 $\delta(t)\in\{0,1,2\}$ 是对于假设 $H_0(t)$ 的决策规则. 令 $R_k=\{t\in T:\delta(t)=k\}$ 表示接受假设 k 的个数. $V_{jk}=\{t\in T:z(t)=j,\delta(t)=k\}$ 表示假设 j 成立且接受假设 k 的个数($k,j=0,1,2$). 表 1 总结了方向性多重假设检验的可能结果.

表 1 方向性多重假设检验的结果

Table 1 Results of directional multiple hypothesis testing

	声明为原假设($\delta(t)=0$)	声明为备择假设 1($\delta(t)=1$)	声明为备择假设 2($\delta(t)=2$)	合计
原假设($z(t)=0$)	V_{00}	V_{01}(第Ⅰ错误)	V_{01}(第Ⅰ错误)	T_0
备择假设 1($z(t)=1$)	V_{10}(第Ⅱ错误)	V_{11}	V_{12}(第Ⅲ错误)	T_1
备择假设 2($z(t)=2$)	V_{20}(第Ⅱ错误)	V_{21}(第Ⅲ错误)	V_{22}	T_2
合计	R_0	R_1	R_2	T

设 $\mathcal{L}(\cdot)$ 是在时间范围上的勒贝格测度,令

$$\begin{cases}\mathcal{L}(N_1)=\mathcal{L}(V_{01})+\mathcal{L}(V_{02}),\\ \mathcal{L}(N_2)=\mathcal{L}(V_{10})+\mathcal{L}(V_{20}),\\ \mathcal{L}(N_3)=\mathcal{L}(V_{12})+\mathcal{L}(V_{21}).\end{cases}$$

式中,$\mathcal{L}(N_1)$和$\mathcal{L}(N_2)$分别为犯第Ⅰ类错误和第Ⅱ类错误的勒贝格测度. $\mathcal{L}(N_3)$为犯第Ⅲ类错误的勒贝格测度,是一种方向性错误. 我们给出针对第Ⅰ类和第Ⅲ类错误的错误发现率(FDR)定义,记为 FDR_{I} 和 FDR_{III},同时给出对应的边际错误发现率定义,记为 $mFDR_{\mathrm{I}}$ 和 $mFDR_{\mathrm{III}}$. 设 $a\vee b=\max(a,b)$,FDR 和 $mFDR$ 的具体定义如下:

$$FDR_{\mathrm{I}}=\boldsymbol{E}\left\{\frac{\mathcal{L}(N_1)}{\mathcal{L}(R_1\cup R_2)\vee 1}\right\}\text{ 和 }mFDR_{\mathrm{I}}=\frac{\boldsymbol{E}\{\mathcal{L}(N_1)\}}{\boldsymbol{E}\{\mathcal{L}(R_1\cup R_2)\}}.$$

$$FDR_{\mathrm{III}}=\boldsymbol{E}\left\{\frac{\mathcal{L}(N_3)}{\mathcal{L}(R_1\cup R_2)\vee 1}\right\}\text{ 和 }mFDR_{\mathrm{III}}=\frac{\boldsymbol{E}\{\mathcal{L}(N_3)\}}{\boldsymbol{E}\mathcal{L}(R_1\cup R_2)}.$$

错误未发现率($FNDR$)和边际错误未发现率($mFNDR$)的定义如下:

$$FNDR=\boldsymbol{E}\left\{\frac{\mathcal{L}(N_2)}{\mathcal{L}(R_0)\vee 1}\right\}\text{ 和 }mFNDR=\frac{\boldsymbol{E}\{\mathcal{L}(N_2)\}}{\boldsymbol{E}\mathcal{L}(R_0)}.$$

从定义中可以看出它们与第Ⅱ类错误有关.

在应用过程中,我们通常会根据实际要求将某种错误发现率或者边际错误发现率控制在某一水平 α 下,从而控制某一类错误发生的概率,得到合理、有效的检验结果.

1.1.2 逐步向下检验过程

基于上述方向性多重假设检验和错误发现率的理论,我们的目标是给出在任一时间点的最优决策规则.

假定模型(1)对应的观测数据为$\{(Y_{1i},t_{1i}):i=1,\cdots,m_1\}$和$\{(Y_{2i},t_{2i}):i=1,\cdots,m_2\}$,我们考虑损失函数

$$L(\delta,z;\lambda)=\lambda_1\mathcal{L}(N_1)+\lambda_2\mathcal{L}(N_2)+\lambda_3\mathcal{L}(N_3),\tag{3}$$

式中,λ_1,λ_2 和 λ_3 是相对成本.

定理 1[11] 设 D 是包含所有$\{(Y_{1i},t_{1i}):i=1,\cdots,m_1\}$和$\{(Y_{2i},t_{2i}):i=1,\cdots,m_2\}$的数据集. 在损失函数(3)中假定 $\lambda_1=\lambda$ 且 $\lambda_2=\lambda_3=1$,则最优决策规则

$$\delta^{(\mathrm{I})}=\{\delta^{(\mathrm{I})}(t):t\in T\}=\mathrm{argmin}_{\delta}E\{L(\delta,z;\lambda)\mid D\}$$

为

$$\delta^{(\mathrm{I})}(t)=\begin{cases}2, & \text{如果}\dfrac{P(z(t)=2\mid D)}{P(z(t)=0\mid D)}\text{且 }P(z(t)=2\mid D)>P(z(t)=1\mid D),\\ 1, & \text{如果}\dfrac{P(z(t)=1\mid D)}{P(z(t)=0\mid D)}\text{且 }P(z(t)=2\mid D)\leqslant P(z(t)=1\mid D),\\ 0, & \text{其他.}\end{cases}$$

定理 1 给出了当 $\lambda_1=\lambda$ 且 $\lambda_2=\lambda_3=1$ 时,使得 $E\{L(\delta,z;\lambda)\mid D\}$ 达到最小值的最优决策规则. 但在实际应用中,我们通常给定某类错误发现率的水平 α,而不是给定具体的损失函数. 定理 2 证明了当给定边际错误发现率 $mFDR_{\mathrm{I}}$ 的水平 α 时,如何搜寻最优的决策规则.

定理 2[11] 设 $\mathcal{A}=\{\delta^{(\mathrm{I})}(t):\lambda>0\}$ 是由定理 1 得到的决策规则 $\delta^{(\mathrm{I})}$ 的集合. 给定边际错误发现率 $mFDR_{\mathrm{I}}$ 的水平为 α,设 $\delta=\{\delta(t):t\in T\}$ 是满足 $mFDR_{\mathrm{I}}\{\delta\}<\alpha$ 的任一决策规则,存在 λ 使得 $\delta^{(\mathrm{I})}\in\mathcal{A}$ 优于 δ,即

$$mFDR_{\mathrm{I}}\{\delta^{(\mathrm{I})}\}\leqslant mFDR_{\mathrm{I}}\{\delta\}\leqslant\alpha\tag{4}$$

和

$$mFNDR\{\delta^{(\mathrm{I})}\}\leqslant mFNDR\{\delta\}.\tag{5}$$

定理 2 证明了在控制 $mFDR_{\mathrm{I}}<\alpha$ 时得到的最优决策规则属于集合 $\mathcal{A}$,即只需在集合 $\mathcal{A}$ 中搜寻最优决策规则,而不是搜寻所有的决策规则. 这大大节约了搜寻最优决策规则的时间. 下面给出了如何在集合 $\mathcal{A}$

中搜寻给定 $mFDR_{\mathrm{I}}<\alpha$ 时的最优决策规则过程.

由于观测到的两个函数型曲线在不同的时间点采样,因此假定区间 $T=[0,1]$,并划分成 M 个等长的子区间$(s_{i-1},s_i]$,其中 $s_0=0$ 和 $s_i=s_{i-1}+1/M,i=1,\cdots,M-1$,选择子区间 $(s_{i-1},s_i]$的中点记为 $t_i^*,i=1,\cdots,M$. 逐步向下检验过程如下:

令

$$\lambda^*=\inf\{\lambda:\widehat{mFDR_{\mathrm{I}}}(\lambda)\leqslant\alpha\},$$

$$\delta^{(\mathrm{I})}(t)=\sum_{i=1}^{M}I(s_{i-1}\leqslant s_i)\delta^{(\mathrm{I})}(t_i^*),$$

有

$$\delta^{(\mathrm{I})}(t_i^*)=\begin{cases}2, & \text{如果}\dfrac{S_2(t_i^*)}{S_0(t_i^*)}>\lambda^*\text{且}\ S_2(t_i^*)>S_1(t_i^*),\\ 1, & \text{如果}\dfrac{S_1(t_i^*)}{S_0(t_i^*)}>\lambda^*\text{且}\ S_2(t_i^*)\leqslant S_1(t_i^*),\\ 0, & \text{其他.}\end{cases}\tag{6}$$

式中,

$$S_k(t)=P(z(t)=k|D),k=0,1,2,$$

$$\widehat{mFDR_{\mathrm{I}}}(\lambda)=\frac{1}{r}\sum_{i=1}^{M}S_1(t_i^*)I(\delta(t_i^*)\neq 0).$$

对于逐步向下检验过程中的未知量 $S_k(t)=P(z(t)=k|\mathcal{D})\ (k=0,1,2)$,可以使用高斯过程回归模型来估计[12].

1.2　信息熵

定义 1　设 X 是一个概率系统中的随机事件,在某一概率空间中有 n 个事件$\{x_i,i=1,\cdots,n\}$,对于 $1\leqslant i\leqslant n$,事件 x_i 发生的概率为 p_i,$\log\dfrac{1}{p_i}$是事件 x_i 发生的信息量. 因此,称$\sum\limits_{i=1}^{n}p_i\log\dfrac{1}{p_i}$为事件 X 的信息熵,记为 $H(X)$.

信息熵描述了随机事件 X 的平均不确定性,即所有可能发生事件所带来的信息量的期望. 信息熵反应了系统不规则程度的度量. 当系统越复杂,出现不同情况的种类越多,那么它的信息熵是较大的,反之则相反. 信息熵的性质和实际意义,为下一节平行度统计量的提出奠定了良好的理论基础.

1.3　基于方向性多重假设检验和信息熵的平行度统计量

1.3.1　平行度统计量

函数型曲线形态会随时间的变化而发生改变. 不同函数型曲线在不同的时间区域上,都可能会有不同的形态差异. 因此在描述函数型曲线时,仅通过欧式距离无法准确地概括两个函数型曲线之间的关系. 有些函数型曲线的距离很接近,但在形态上差异很大. 我们由此受到启发提出一种新的平行度统计量,用来刻画函数型曲线之间的形态差异.

对于模型(1):$\begin{cases}Y_1(t)=\mu(t)+\mu_d(t)+\varepsilon_1(t)\\ Y_2(t)=\mu(t)+\varepsilon_2(t).\end{cases}$,1.1.2 节给出了在 $mFDR_I\{\delta\}<\alpha$ 时的逐步向下检验过程(6),可知 $\delta^{(\mathrm{I})}(t)$是最优检验过程中给定时间点 t 的决策结果,取值为 0、1 和 2. 给定 Δ 时,我们计算两个函数型曲线决策结果$\{\delta^{(\mathrm{I})}(t):t\in T\}$不同取值在时间范围 T 内的占比,分别记为 $\Gamma_0(\Delta)$、$\Gamma_1(\Delta)$和 $\Gamma_2(\Delta)$:

$$\begin{cases}\Gamma_0(\Delta)=\dfrac{\sum_{i=1}^{m}I\{\delta^{(\mathrm{I})}(t)=0\}}{m},\Gamma_1(\Delta)=\dfrac{\sum_{i=1}^{m}I\{\delta^{(\mathrm{I})}(t)=1\}}{m},\Gamma_2(\Delta)=\dfrac{\sum_{i=1}^{m}I\{\delta^{(\mathrm{I})}(t)=2\}}{m},\\ \Gamma_0(\Delta)+\Gamma_1(\Delta)+\Gamma_2(\Delta)=1.\end{cases}\tag{7}$$

这 3 个统计量描述了在给定 Δ 下两条函数型曲线的位置关系. 当 $\Gamma_0(\Delta)$取值较大时,表示在大部分时间点上 $Y_1(t)$和 $Y_2(t)$间的差异分布在区间$[-\Delta,\Delta]$,此时可认为两条曲线无大的形态差异;当 $\Gamma_1(\Delta)$取值较大时,表示 $Y_1(t)$大部分时间点分布在 $Y_2(t)$的下方,且两者之间的差异大于Δ;当 $\Gamma_2(\Delta)$取值较大时,

表示 $Y_1(t)$ 大部分时间点分布在 $Y_2(t)$ 的上方,且两者之间的差异大于 Δ. 其中,$\Gamma_0(\Delta)$ 随着 Δ 的变化单调递增,$\Gamma_1(\Delta)$ 和 $\Gamma_2(\Delta)$ 随着 Δ 的变化单调递减.

给定 $0<\Delta_1<\Delta_2<\cdots<\Delta_P$,其中 Δ_P 表示两条曲线差异的最大值. 我们可得到表 2 的数据.

表 2　不同 Δ 下对应的 $\Gamma_i(\Delta)(i=0,1,2)$

Table 2　Corresponding $\Gamma_i(\Delta)(i=0,1,2)$ at different Δ

Δ	Δ_1	Δ_2	$\cdots$	Δ_{P-1}	Δ_P
Γ_0	Γ_{01}	Γ_{02}	$\cdots$	$\Gamma_{0(P-1)}$	Γ_{0P}
Γ_1	Γ_{11}	Γ_{12}	$\cdots$	$\Gamma_{1(P-1)}$	Γ_{1P}
Γ_2	Γ_{21}	Γ_{22}	$\cdots$	$\Gamma_{2(P-1)}$	Γ_{2P}

基于表 2,对所分布的区间重新划分为 $[-\Delta_P,-\Delta_{P-1})$、$\cdots$、$[-\Delta_2,-\Delta_1)$、$[-\Delta_1,\Delta_1)$、$[\Delta_1,\Delta_2)$、$\cdots$、$[\Delta_{P-1},\Delta_P)$,并统计落在各个区间上的频率,见表 3:

表 3　当 $H_0(t):\mu_d(t)\in D$ 时,多重假设检验 Γ_0 的结果

Table 3　Results of multiple hypothesis testing Γ_0 when $H_0(t):\mu_d(t)\in D$

D	$[-\Delta_P,-\Delta_{P-1})$	$\cdots$	$[-\Delta_2,-\Delta_1)$	$[-\Delta_1,\Delta_1)$	$[\Delta_1,\Delta_2)$	$\cdots$	$[\Delta_{P-1},-\Delta_P)$
$\Delta\Gamma$	$\Gamma_{1(P-1)}-\Gamma_{1P}$	$\cdots$	$\Gamma_{11}-\Gamma_{12}$	Γ_{01}	$\Gamma_{21}-\Gamma_{22}$	$\cdots$	$\Gamma_{2(P-1)}-\Gamma_{2P}$

其中,D 表示两条曲线的差异区间,$\Delta\Gamma$ 表示 $\widehat{\mu_d}(t)$ 落在每一区间上的频率. 通过比较各个区间的 $\Delta\Gamma$ 值,我们可以大致得到差异曲线 $\widehat{\mu_d}(t)$ 的分布情况.

差异曲线可视为信息源发出的信号,信号系统越复杂,信息熵的值越大,反之则相反. 对于两条曲线而言,若差异曲线都集中在一个区间内,则可认为两条曲线的形态是相似的,我们可看作是平行的,此时信息熵最小;若差异曲线散布在各个区间上,则两条曲线的形态差异大、相似度低,此时可认为不平行. 基于上述思想,我们提出了基于信息熵和方向性多重假设检验的平行性统计量 $H(Y_1(t),Y_2(t),D)$:

$$\begin{cases} H(Y_1(t),Y_2(t),D)\equiv H(\widehat{\mu_d}(t),D)=-\sum\limits_{p=1}^{2P-1}\Delta\Gamma_p\log\Delta\Gamma_p, \\ \Delta\Gamma_1+\Delta\Gamma_2+\cdots+\Delta\Gamma_P=1. \end{cases} \tag{8}$$

式中,$\Delta\Gamma_p$ 为表 3 中区间划分 D 的第 p 个区间所对应的 $\Delta\Gamma$ 值,$p=1,2,\cdots,2P-1$. 平行度统计量更好地描述了两曲线之间的形态差异. 当统计量越大时表示平行度越低,形态差异越大;当统计量越小时表示平行度越高,形态差异越小.

值得我们关注的是对区间的划分. 往往会依据有关专家的建议或者我们感兴趣的差异大小选取合适的间隔,将区间等距离地划分. 显而易见的是,当间隔划分得越细时表示我们对平行度的要求越高,间隔划分得越粗时表示对平行度的要求越低,而过分地细划分或者粗划分对平行度的度量显然是没有意义的.

1.4　基于平行度和欧式距离的凝聚式层次聚类算法

对函数型数据集 $\{y_i(t),i=1,\cdots,n\}$ 进行聚类时,常用的函数型非参数方法主要基于曲线间的距离. 但在实际中存在着属于不同类别的两条函数型曲线距离很接近,但形态差异很大. 因此仅仅依靠距离对函数型曲线进行聚类显然是不够的. 结合欧式距离和平行度统计量,本文提出了对函数型曲线聚类的一个新的接近度指标,它更好地概括了曲线之间的距离差距和形态差异.

对任意两条曲线 $y_i(t)$ 与 $y_j(t)$ 来说,$d^1_{i,j}$ 表示它们之间的欧式距离,进行归一化处理得到. 具体过程如下:

$$d^1_{i,j}=\sqrt{\sum_{k=1}^{m}(y_i(t_k)-y_j(t_k))^2},$$

归一化处理后,有

$$d^{1'}_{i,j}=\frac{d^1_{i,j}-\min\{d^1_{k,q}\}}{\max\{d^1_{k,q}\}-\min\{d^1_{k,q}\}},\quad k\neq q.$$

同理,计算得到归一化后的平行度统计量:

$$d^{2'}_{i,j}=\frac{H(y_i(t),y_j(t),D)-\min\{H(y_k(t),y_q(t),D)\}}{\max\{H(y_k(t),y_q(t)),D\}-\min\{H(y_i(j),y_q(t)),D\}},\quad k\neq q.$$

我们引入 $p(0\leqslant p\leqslant 1)$ 表示权重,从而构造了新的指标来反映两个函数型曲线之间的接近度:

$$\mathrm{dist}(y_i(t),y_j(t),p,D)=p\times d_{i,j}^{1'}+(1-p)\times d_{i,j}^{2'},\tag{9}$$

式中,$i\neq j,1\leqslant i,j\leqslant n$.

式(9)中定义的新指标同时考虑了平行度和距离两方面的信息. 而 p 作为一个权重,反映了距离和平行度对新指标的贡献程度. 当 p 越大时距离在聚类中起主导作用. 当 $p=1$ 时,新指标就是我们所熟悉的欧式距离. 反之,当 p 越小时平行度在聚类中起主导作用.

基于新的聚类指标,我们对于函数型数据提出了一种新的凝聚式层次聚类算法. 设有 n 个待聚类的函数型数据样本,凝聚式层次聚类算法的步骤如表 4 所示:

表 4　新的凝聚式层次聚类算法

Table 4　New cohesive hierarchical clustering algorithm

算法 1　凝聚式层次聚类算法
1.初始化:数据集中的每个对象生成一个类,得到类列表 $C=\{c_1(t),c_2(t),\cdots,c_n(t)\}$
(a)每个簇只包含一个数据对象:$c_i(t)=\{y_i(t)\}$;
2.重复如下步骤,直到 C 中只有一个簇:
(a)计算种类之间的相似度,从 C 中找到相似度最接近的两个类 $c_i(t)$ 和 $c_j(t)$
(b)合并 $c_i(t)$ 和 $c_j(t)$,生成新类 $c_{(i+j)}(t)$;
(c)从 C 中删除类 $c_i(t)$ 和类 $c_j(t)$,添加新类 $c_{(i+j)}(t)$.

在聚类的过程中,当平行度相似时,距离较近的两个类首先聚到一起,而当距离相似时,平行度更接近的两个类首先聚到一起. 另一方面,p 调节了距离和平行度在接近度计算过程中所占的比重. 当区间划分给定时,p 的选择可以基于如下准则函数:

$$Op(p)=\frac{\sum_{i=1}^{n}d_{i,k_i}^{1}(p)}{\max\sum_{i=1}^{n}d_{i,k_i}^{1}(p)}+\frac{\sum_{i=1}^{n}H_{i,k_i}(p)}{\max\sum_{i=1}^{n}H_{i,k_i}(p)},\tag{10}$$

式中,$i\in\{1,\cdots,n\}$ 表示第 i 个样本,$k_i\in\{1,\cdots,K\}$ 表示第 i 个样本聚类后所属的类,$d_{i,k_i}^{1}(p)$ 表示第 i 个样本与所在类的类中心之间的欧式距离,$H_{i,k_i}(p)$ 表示第 i 个样本与所在类的类中心的平行度统计量.

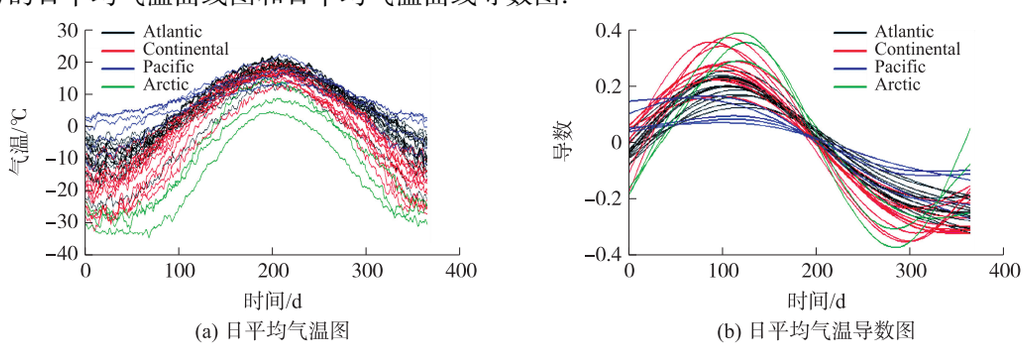
式(10)由两个部分组成,第一部分计算了样本数据集聚类后所有样本与对应所在类的类中心的距离之和,第二部分计算了所有样本与对应所在类的类中心的平行度之和,它们刻画了类内的差异. 最优的 p 就是 O_p 的极小值点.

2　实证分析

为了检验本文提出的基于欧式距离和平行度的新的凝聚式层次聚类算法的有效性,我们考虑了加拿大天气、身高、酵母基因和水分含量 4 个函数型数据集. 其中,加拿大天气和身高数据集从 R 软件的 fda 包中获取,酵母基因和水分含量数据集从文[4]中获取. 用准确率作为评价指标.

2.1　加拿大天气聚类分析

加拿大天气数据集中有从 1960 年到 1994 年间加拿大 35 个气象站 365 天的日平均气温,以及 35 个气象站所处的 4 个不同气候区(分别是 Atlantic、Continental、Pacific 和 Arctic). 图 1 为 35 个气象站原始(未光滑)的日平均气温曲线图和日平均气温曲线导数图.



(a) 日平均气温图　　(b) 日平均气温导数图

图 1　加拿大天气日平均气温(左图)和日平均气温导数图(右图)

Fig. 1　Canadian weather daily average temperature(left graph)and daily average temperature derivative graph(right graph)

从图 1 可以看到 Pacific 气候区的气象站常年气温较高且变化平缓,而位于 Arctic 气候区的气象站一年绝大多数时间段处于零下摄氏度,且气温变化幅度大. Atlantic 和 Continental 气候区的气象站无论是气温曲线还是气温导数曲线均介于前两者之间. 由此,我们可以发现不同气候区气象站的日平均气温曲线不仅存在着距离差异,在曲线的形状上也有一定的不同.

利用第 1.3 节所提出的方法,计算各条差异曲线的平行度统计量和欧式距离. 我们发现当 $p \in \{0.4, 0.5\}$ 且 $\Delta_{gap}=2$ 时,准确率达到最大,为 0.8.

表 5 罗列了其它 4 种聚类方法在不同初始化方法下对加拿大天气数据集聚类的准确率. 分别是 Classical K-means(传统 k-means)方法、funHDDC[13]、funclust 方法[7](标记为 funPEM)和 Zambom 等[4]提出的基于平行度检验统计量和均值相等检验统计量的度量方法,该度量方法改进了 k-means 方法,被命名为 Tested-based K-means. 从表 5 我们可以清楚地看到,本文提出的聚类方法的准确率明显高于其他方法.

表 5　其它聚类方法下加拿大天气的准确率

Table 5　Accuracy of Canadian weather under other clustering methods

聚类方法	Classical K-means			Test-based K-means				funHDDC			funFEM		
	随机	K-means	层次聚类	随机	K-means	层次聚类	K-means++	随机	K-means	层次聚类	随机	K-means	层次聚类
准确率	0.64	0.59	0.59	0.60	0.63	0.62	0.64	0.55	0.59	0.63	0.40	0.67	0.71

我们也计算了给定 Δ_{gap} 下不同 p 所对应的准则 O_p,发现在 $\Delta_{gap}=2$,$p=0.5$ 所对应的准则 O_p 最小,取值为 1.573. 此时,各个样本到所在类中心的欧式距离和平行度统计量之和最小,即类内差异最小,我们有理由认为此时的聚类效果最好.

2.2　身高的聚类分析

身高数据集中包含了 39 个男生和 54 个女生的身高数据,以及他们在 1 岁到 18 岁的 31 个被记录的时间点. 图 2 为 93 个孩子的身高图. 从图中我们可以看到这是典型的单调递增函数型数据. 女生比男生更早结束身高生长期,且最终大部分女生的身高低于男生身高. 我们可以发现,两类曲线有一定的距离差距和形态差异.

通过计算,我们发现在 $\Delta_{gap}=3$ 时,$p=0.8$ 所对应的准则函数 O_p 最小,取值为 1.6477,准确率为 0.72,聚类效果较好,优于表 6 中 Classical K-means 方法和 funclust 方法的聚类效果.

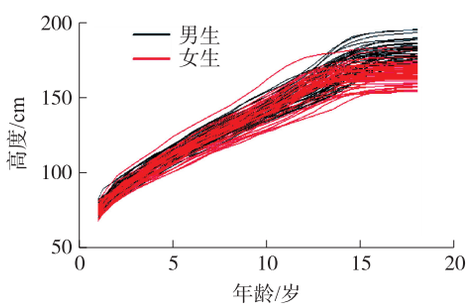


图 2　身高图

Fig. 2　Graph of height date

表 6　其它聚类方法下身高的准确率

Table 6　Accuracy of height under other clustering methods

聚类方法	Classical K-means			Test-based K-means				funHDDC			funFEM		
	随机	K-means	层次聚类	随机	K-means	层次聚类	K-means++	随机	K-means	层次聚类	随机	K-means	层次聚类
准确率	0.69	0.67	0.67	0.74	0.74	0.74	0.74	0.88	0.88	0.88	0.68	0.68	0.68

2.3　酵母基因的聚类分析

图 3 是 78 个酵母基因的对数基因表达比率. 表达比率每 7 min 被记录一次,共有 18 个时间点. 生物学家表示此数据集中有 5 个类,其中样本 1 到 13 是 G1 类,样本 14 到 52 是 S 类,样本 53 到 60 是 S/G2 类,样本 61 到 67 是 G2/M 类,样本 68 到 78 是 M/G1. 从图 3 可以看出,不同类之间存在相位的差异,而同一类里面的样本存在振幅的差异. 相比于加拿大天气和身高两个数据集,酵母基因数据集无论是不同类之间还是个体之间的差异都更复杂.

通过计算,我们发现在 $\Delta_{gap}=0.2$ 时,$p=1$ 所对应的准则函数 O_p 取值最小,取值为 1.523 8,此时准确率为 0.538. 此

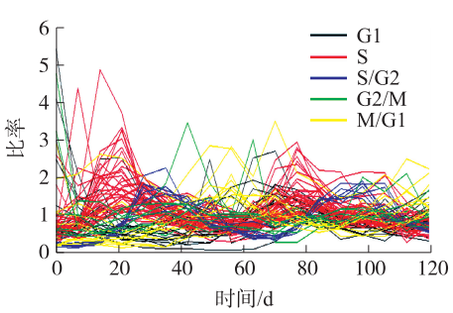


图 3　酵母基因图

Fig. 3　Graph of yeast gene date

时,本文所提出的聚类方法优于表 7 中其它聚类方法.

表 7　其它聚类方法下酵母基因的准确率

Table 7　Accuracy of yeast genes under other clustering methods

聚类方法	Classical *K*-means			Test-based *K*-means				funHDDC			funFEM		
	随机	*K*-means	层次聚类	随机	*K*-means	层次聚类	*K*-means++	随机	*K*-means	层次聚类	随机	*K*-means	层次聚类
准确率	0.49	0.51	0.50	0.51	0.53	0.50	0.49	0.42	0.45	0.44	0.41	0.44	0.44

2.4　水分含量的聚类分析

脂肪光谱数据集由 100 个小麦样品的近红外反射光谱组成,近红外反射光谱在 1 100 nm 到 2 500 nm 之间每 3 nm 被测量一次. 水分含量是与近红外反射光谱有关的相应变量. 图 4 为 100 个小麦样品所对应的水分含量图,人为地将 100 个数据分为两类,其中样本 1 到 41 为第一类,样本 42 到 100 为第二类. 从图 4 中,我们可以明显地看到两类曲线的变化趋势完全相同,但第一类曲线的方差大于第二类曲线的方差,即第一类曲线的分布范围更大,第二类曲线分布地更紧凑. 下面考察对于类方差不同的函数型数据集,本文提出的方法是否具有一定优势.

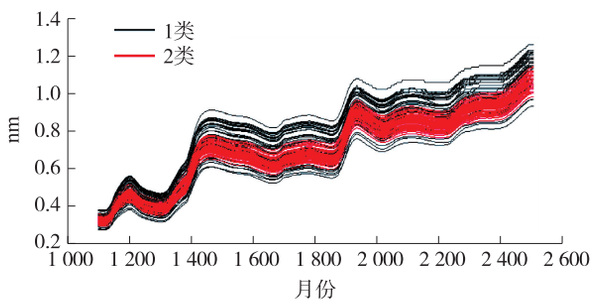


图 4　脂肪光谱数据图

Fig. 4　Graph of fat spectrum date

通过计算我们发现在 $\Delta_{\text{gap}}=0.03$ 时,$p=0.4$ 所对应的准则函数 O_p 最小,取值为 1.403 6,对应的准确率为 0.81,此时的聚类效果优于表 8 中的 Classical *K*-means 聚类方法、funHDDC 方法 和 funclust 方法.

表 8　其它聚类方法下脂肪光谱数据的准确率

Table 8　Accuracy of fat spectral data under other clustering methods

聚类方法	Classical *K*-means			Test-based *K*-means				funHDDC			funFEM		
	随机	*K*-means	层次聚类	随机	*K*-means	层次聚类	*K*-means++	随机	*K*-means	层次聚类	随机	*K*-means	层次聚类
准确率	0.78	0.78	0.78	0.83	0.83	0.83	0.83	0.78	0.78	0.78	0.78	0.78	0.78

3　结论

本文针对函数型数据提出了基于欧式距离和平行度的接近度指标,提出了一种新的凝聚式层次聚类算法,改进了单独使用距离聚类带来的局限性,而且充分考虑了函数型曲线的形状. 与其它方法相比,该方法有下述的优点:

(1)基于原始数据的函数型数据聚类方法往往是通过距离展开的. 但在实际应用中,我们常常会碰到属于不同类别的函数型曲线,它们之间的距离很接近,但形态差异很大. 若仍使用原来的聚类方法,可能无法实现准确聚类. 本文基于欧式距离和平行度提出了一个新的接近度指标,它综合平行度和距离的信息,更好地概括了两个函数型曲线之间的差异,最终优化了凝聚式层次聚类方法.

(2)除了考虑平行度和距离的信息,本文在接近度指标的计算过程中,通过 p 反映两类信息对接近度的贡献程度. 当属于不同类别的函数型曲线的差异主要体现在距离上时,我们可以适当增加距离的比重. 反之,当属于不同类别的函数型曲线的差异主要体现在形态变化上时,我们也可以适当增加平行度统计量的比重. 针对不同特征的函数型数据,依据准则 O_p 调整权重得到了较好的聚类效果.

但是,该方法还是存在着一些不足之处:

(1)该聚类方法的有效性在一定程度上依赖于平行度统计量,而平行度统计量描述形态差异的效果依赖于对区间的划分. 当间隔划分得越细时表示对平行度的要求越高,间隔划分得越粗时表示对平行度的要求越低. 因此,过分地细划分或者粗划分都会对聚类结果造成一定的影响.

(2)该聚类方法只考虑了类与类间具有距离差距和形态差异的函数型数据集,但现实生活中的函数型数据集是复杂的,不可能只存在距离差距或者形态差异,还存在方差、振幅等其它方面的差异,因此针对 4 种不同类型的函数型数据集,聚类效果并未一致优于其它聚类方法.

综合上述以及实证分析结果,我们可知当函数型数据集中存在明显的距离差距和形态差异时,本文所提出的聚类方法具有一定优势,能得到较好的聚类结果.

[参考文献]

[1] BOULLÉ M, GUIGOURÈS R, ROSSI F. Nonparametric hierarchical clustering of functional data [M]. Chapman: Springer, 2014.

[2] IEVA F, PAGANONI A M, PIGOLI D, et al. Multivariate functional clustering for the morphological analysis of electrocardiograph curves[J]. Journal of the royal statistical society, 2013, 62(3): 401-418.

[3] TOKUSHIGE S, YADOHISA H, INADA K. Crisp and fuzzy *k*-means clustering algorithms for multivariate functional data[J]. Computational statistics, 2007, 22: 1-16.

[4] ZAMBOM A Z, COLLAZOS J A A. Functional data clustering via hypothesis testing *k*-means[J]. Computational statistics, 2019, 34(2): 527-549.

[5] JAMES G M, SUGAR C A. Clustering for sparsely sampled functional data[J]. Journal of the american statistical association, 2003, 98(462): 397-408.

[6] CHIOU J M, LI P L. Functional clustering and identifying substructures of longitudinal data[J]. Journal of the royal statistical society, 2007, 69(4): 679-699.

[7] JACQUES J, PREDA C. Functional data clustering: a survey[J]. Advances in data analysis and classification, 2013, 8(3): 231-255.

[8] TARPEY T, KINATEDER K K J. Clustering functional data[J]. Journal of classification, 2003, 20(1): 93-114.

[9] HEARD N A, HOLMES C C, STEPHENS D A. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of bayesian hierarchical clustering of curves[J]. Journal of the American Statistical Association, 2006, 101(473): 18-29.

[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] XU P, LEE Y, SHI J Q. Automatic detection of significant areas for functional data with directional error control[J]. Statistics in medicine, 2018, 38(3): 376-397.

[12] SHI J Q, CHOI T. Gaussian process regression analysis for functional data[M]. New York: Chapman and Hall, 2011.

[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.

[责任编辑:陆炳新]