ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 1,Jan. 2021,pp. 45－52
DOI:10. 16337/j. 1004-9037. 2021. 01. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种基于张量分解的医学数据缺失模态的补全算法

刘　琚，杜若画，吴　强，何泽鲲，于璐跃

（山东大学信息科学与工程学院，青岛 266237）

摘　要：多模态磁共振影像数据采集过程中会出现不同程度的模态数据缺失，现有的补全方法大多只针对随机缺失，无法较好地恢复条状及块状缺失。针对此问题，本文提出了一种基于多向延迟嵌入的平滑张量补全算法分类框架。首先，对缺失数据进行多向延迟嵌入操作，得到折叠后的张量；然后通过平滑张量CP分解，得到补全的张量；最后利用多向延迟嵌入的逆向操作，得到补全的数据。该算法在BraTS脑胶质瘤影像数据集上进行了高低级别肿瘤分类实验，并与7种基线模型进行了比较。实验结果表明，本文提出方法的平均分类准确率可达91.31%，与传统补齐算法相比具有较好的准确性。

关键词：张量分解；脑肿瘤分类；缺失模态；数据补全

中图分类号：TP39　　**文献标志码**：A

A Complete Algorithm for Missing Modalities of Medical Data Based on Tensor Decomposition

LIU Ju, DU Ruohua, WU Qiang, HE Zekun, YU Luyue

(School of Information Science and Engineering, Shandong University, Qingdao 266237, China)

Abstract: In the process of multi-modality magnetic resonance image (MRI) data acquisition, there will be different degrees of modality data missing. However, most of the existing completion methods only aim at random missing, which cannot recover strip and block missing. Therefore, this paper proposes a classification framework of smooth tensor completion algorithm based on multi-directional delay embedding. Firstly, the folded tensor is obtained by multi-directional delay embedding of missing data. Then, the completed tensor is obtained by smoothing tensor CP decomposition. Finally, the reverse operation of multi-directional delay embedding is used to obtain the completed data. The algorithm is used to classify high-level and low-level tumors on the BraTS glioma image data set and compared with seven baseline models. The average classification accuracy of the proposed method achieves 91.31%, and experimental results show that the method has better accuracy compared with the traditional complement algorithm.

Key words: tensor factorization; brain tumor classification; missing modality; data completion

引　　言

随着医学技术的发展，磁共振(Magnetic resonance imaging, MRI)技术已广泛应用于临床的疾病诊

基金项目：山东省重点研发计划(2017CXGC1504)资助项目。
收稿日期：2020-11-29；**修订日期**：2021-01-08

断,MRI影像也成为辅助医生进行肿瘤分级的重要参考依据。由于不同模态的MRI影像能够突出肿瘤不同的信息与特点,医生在肿瘤分级时会结合多个模态的MRI影像进行诊断。然而,在实际的MRI影像采集过程中,例如病人头部抖动或缺少某项检查,会导致数据的部分缺失。这些缺失数据在训练模型时被舍弃,造成了数据的浪费,不利于分类结果。为了有效利用这些缺失的数据,减少小样本问题的影响,研究人员提出了许多方法,其中对缺失数据进行补全是有效方法之一[1-3]。传统的缺失补全算法包括:期望最大化算法[4]、奇异值分解算法、邻近算法[5]和矩阵补全[6-7]。但是,这些方法无法有效恢复块状缺失数据,并且对于整个模态的缺失,估算大量缺失值会导致算法性能不稳定。为了应对上述情况,文献[8]提出了基于多向延迟嵌入的Tucker分解算法,考虑张量嵌入空间的低秩模型,但该方法并未考虑到提取特征和补全张量过程中可能出现的噪声。

本文提出了一种基于多向延迟嵌入的平滑张量补全算法(Multi-way delay-embedding transform smooth PAPAFAC tensor completion, MDT-SPC)。通过对输入的不完整张量使用多向延迟嵌入转换,得到不完整的Hankel张量,然后利用平滑CP分解[9]来恢复不完整张量的缺失部分,最后使用多向延迟嵌入的逆过程,得到最终的补全后的张量。该方法能够有效针对数据呈条状或块缺失的情况,并且当提取特征和补全张量的过程中出现噪声时,平滑约束有助于消除计算过程中产生的非平滑项。实验部分采用BraTS2017作为训练和测试数据集,并与多种传统方法和基于多向延迟嵌入的Tucker分解算法进行比较。实验结果表明本文提出的脑胶质瘤分类方法可以获得更高的识别准确率。

1 基于多向延迟嵌入的Tucker分解算法

基于多向延迟嵌入的Tucker分解算法由3步组成:(1)使用多向延迟嵌入转换[8],输入待补全的张量,输出不完整的Hankel张量;(2)通过Tucker低秩分解[10]来恢复不完整张量的缺失部分;(3)使用多向延迟嵌入的逆过程[8],得到最终的补全后的张量。

1.1 多向延迟嵌入及其逆变换

假设向量为$\boldsymbol{v}=(v_1,v_2,\cdots,v_L)^{\mathrm{T}}\in\mathbf{R}^L$,则$\boldsymbol{v}$对应的标准延迟嵌入变换为

$$H_\tau(\boldsymbol{v})=\begin{pmatrix} v_1 & v_2 & \cdots & v_{L-\tau+1} \\ v_2 & v_3 & \cdots & v_{L-\tau+2} \\ \vdots & \vdots & \ddots & \vdots \\ v_\tau & v_{\tau+1} & \cdots & v_L \end{pmatrix}\in\mathbf{R}^{\tau\times(L-\tau+1)} \tag{1}$$

式中τ为延迟嵌入变换的重复次数。标准延迟嵌入变换$H_\tau(\boldsymbol{v})=\mathrm{fold}_{(L,\tau)}(\boldsymbol{S}\boldsymbol{v})$,正向变换的过程是复制和折叠,而逆变换可以分解为矢量化运算和Moore-Penrose伪逆$\boldsymbol{S}^\dagger:=(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})^{-1}\boldsymbol{S}^{\mathrm{T}}$。因此,Hankel矩阵$\boldsymbol{V}_{\mathrm{H}}$的逆延迟嵌入变换为$\boldsymbol{H}_\tau^{-1}(\boldsymbol{V}_{\mathrm{H}})=\boldsymbol{S}^\dagger\,\mathrm{vec}(\boldsymbol{V}_{\mathrm{H}})$。

多向延迟嵌入变换可以表示为$H_\tau(\boldsymbol{T})=\mathrm{fold}_{(I,\tau)}\boldsymbol{T}\times_1\boldsymbol{S}_1\times_2\boldsymbol{S}_2\cdots\times_N\boldsymbol{S}_N$,符号"$\times_1$""$\times_2$","$\times_N$"分别代表张量与矩阵在第$N$阶上的乘法操作,最后,张量的逆多向延迟嵌入变换为$H_\tau^{-1}(\boldsymbol{T})=\mathrm{unfold}_{(1,\tau)}(\boldsymbol{X})\times_1\boldsymbol{S}_1^\phi\cdots\times_N\boldsymbol{S}_N^\phi$,其中$\mathrm{unfold}_{(1,\tau)}=\mathrm{fold}_{(1,\tau)}^{-1}$。

1.2 增秩Tucker分解

$\boldsymbol{T}\in\mathbf{R}^{I_1\times\cdots\times I_N}$和$\boldsymbol{Q}\in\{0,1\}^{I_1\times\cdots\times I_N}$分别为不完整张量和掩模张量,经过多向延迟嵌入转换之后得到$\boldsymbol{T}_{\mathrm{H}}=H(\boldsymbol{T})\in\mathbf{R}^{J_1\times\cdots\times J_M}$和$\boldsymbol{Q}_{\mathrm{H}}=H(\boldsymbol{Q})\in\{0,1\}^{J_1\times\cdots\times J_M}$,其中$M\geqslant N$。基于Tucker分解模型得到$\boldsymbol{T}_{\mathrm{H}}$的低秩近似,采用交替最小二乘求解,并引入辅助变量来简化优化过程。

基于Tucker的张量补全方法的一个难点在于如何确定适当的秩$(R_1,\cdots,R_M)$,算法采用一种"秩递增"的方法来确定最佳秩,主要思想是张量应通过比其目标秩更低的秩近似来初始化。步骤如下:首先

对于所有的m初始化$R_m=1$,然后使用$(R_1,\cdots,R_M)$计算获得$\boldsymbol{G}、\left\{\boldsymbol{U}^{(m)}\right\}_{m=1}^{M}$和$\boldsymbol{X}=\boldsymbol{G}\times\{\boldsymbol{U}\}$,最后检查噪声条件$\left\|\boldsymbol{Q}_{\mathrm{H}}\circledast(\boldsymbol{T}_{\mathrm{H}}-\boldsymbol{X})\right\|_{\mathrm{F}}^{2}\leqslant\varepsilon$,其中$\circledast$表示张量的Hadamard乘积。若满足则终止算法,否则选择增加的模式m'和增加的$R_{m'}$,重新计算$\boldsymbol{G}、\left\{\boldsymbol{U}^{(m)}\right\}_{m=1}^{M}$和$\boldsymbol{X}=\boldsymbol{G}\times\{\boldsymbol{U}\}$。

2　基于多向延迟嵌入的平滑张量补全算法

增秩Tucker分解采用了较为新颖的秩逼近策略,但是并未考虑提取特征和补全张量过程中出现的噪声,而平滑特性在图像处理中具有很好的可用性。当计算过程产生噪声时,平滑约束有助于消除计算过程中产生的非平滑项,因此在进行多向延迟嵌入操作后,对折叠数据采用平滑CP分解模型[9]进行张量补全。

2.1　平滑CP分解模型

平滑CP分解的优化问题[9]可表示为

$$\begin{aligned}&\min_{G,U^{(1)},\cdots,U^{(N)}}\frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{Z}\right\|_{\mathrm{F}}^{2}+\sum_{r=1}^{R}\frac{g_r^2}{2}\sum_{n=1}^{N}\rho^{(n)}\left\|\boldsymbol{L}^{(n)}\boldsymbol{u}_r^{(n)}\right\|_p^p\\&\text{s.t.}\quad\boldsymbol{Z}=\sum_{r=1}^{R}g_r\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}\\&\qquad\boldsymbol{X}_{\Omega}=\boldsymbol{T}_{\Omega},\boldsymbol{X}_{\bar{\Omega}}=\boldsymbol{Z}_{\bar{\Omega}},\left\|\boldsymbol{u}_r^{(n)}\right\|_2=1\\&\qquad\forall r\in\{1,\cdots,R\},\forall n\in\{1,\cdots,N\}\end{aligned}\tag{2}$$

式中:$\boldsymbol{X}$为恢复的完整张量,其中的缺失元素由CP分解估计值$\boldsymbol{Z}$来填充;Ω表示$\boldsymbol{T}$的可用元素索引,即未缺失数据的坐标,$\bar{\Omega}$表示$\boldsymbol{T}$缺失元素的索引;$\boldsymbol{\rho}=[\rho^{(1)},\rho^{(2)},\cdots,\rho^{(N)}]^{\mathrm{T}}$表示平滑参数;$p\in\{1,2\}$代表用来选择平滑约束类型的参数;矩阵$\boldsymbol{L}^{(n)}\in\mathbf{R}^{(I_n-1)\times I_n}$代表平滑约束矩阵。

目标函数第一项$\left\|\boldsymbol{X}-\boldsymbol{Z}\right\|_{\mathrm{F}}^{2}$代表观察值$\boldsymbol{T}_{\Omega}$和CP分解模型$\boldsymbol{Z}_{\Omega}$之间的均方误差,因此第一项提供输入张量$\boldsymbol{T}$的CP分解;目标函数第二项是惩罚项,用来确保CP分解的因子矩阵$\boldsymbol{u}_r^{(n)}$是平滑的,有$\left\|\boldsymbol{L}^{(n)}\boldsymbol{u}_r^{(n)}\right\|_p^p=\sum_{i=1}^{I_n-1}\left|\boldsymbol{u}_r^{(n)}(i)-\boldsymbol{u}_r^{(n)}(i+1)\right|^p$,这种非平滑测度的最小化加强了单个因子矩阵$\boldsymbol{u}_r^{(n)}$的平滑性。当$p=1$时,约束项为总变分约束,当$p=2$时约束项变为二次变分约束;$g_r^2$为自适应因子,它取决于平滑约束项的惩罚水平;$R$代表张量CP分解中秩的个数,$N$为张量阶数。

2.2　算法求解

对目标函数使用分层交替最小二乘法求解,以因子矩阵为单元逐一更新,分别处理每个因子矩阵的规范约束为

$$\begin{aligned}&\min_{g,u_r^{(1)},\cdots,u_r^{(N)}}\frac{1}{2}\left\|\boldsymbol{Y}_r-\boldsymbol{Z}_r\right\|_2^{\mathrm{F}}+\frac{g_r^2}{2}\sum_{n=1}^{N}\rho^{(n)}\left\|\boldsymbol{L}^{(n)}\boldsymbol{u}_r^{(n)}\right\|_p^p\\&\text{s.t.}\quad\boldsymbol{Z}_r=g_r\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}\\&\qquad[\boldsymbol{Y}_r]_{\Omega}=\boldsymbol{T}_{\Omega}^{(r)},[\boldsymbol{Y}_r]_{\bar{\Omega}}=[\boldsymbol{Z}_r]_{\bar{\Omega}}\\&\qquad\left\|\boldsymbol{u}_r^{(n)}\right\|_2=1\quad\forall n\in\{1,\cdots,N\}\end{aligned}\tag{3}$$

式中$\boldsymbol{Y}_r=\boldsymbol{X}-\sum_{i\neq r}g_i\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}$,$\boldsymbol{T}_{\Omega}^{(r)}=\boldsymbol{T}_{\Omega}-\left[\sum_{i\neq r}g_i\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}\right]_{\Omega}$,那么这个局部求解问题仅涉及CP分解的第$r$个分量。为了求解上述问题,按以下顺序更新$\boldsymbol{u}_r^{(1)},\boldsymbol{u}_r^{(2)},\cdots,\boldsymbol{u}_r^{(N)},g_r$并重置$\boldsymbol{Y}_r$为

$$\boldsymbol{u}_r^{(n)} \leftarrow \underset{u\in U^{I_n}}{\arg\min}\, F_p^{(r,n)}(\boldsymbol{u}) \tag{4}$$

$$g_r \leftarrow \underset{g}{\arg\min}\, F_p^{(r)}(g) \tag{5}$$

$$[\boldsymbol{Y}_r]_{\bar{\Omega}} \leftarrow [g_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(N)}]_{\bar{\Omega}} \tag{6}$$

式中

$$F_p^{(r)}(g_r) = F_p^{(r,n)}(\boldsymbol{u}_r^{(n)}) = \frac{1}{2}\left\| \boldsymbol{Y}_r - g_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(N)} \right\|_{\mathrm{F}}^2 + \frac{g_r^2}{2}\sum_{n=1}^{N} \rho^{(n)} \left\| L^{(n)} \boldsymbol{u}_r^{(n)} \right\|_p^p \tag{7}$$

使用 $\boldsymbol{u}_k$ 和 $\boldsymbol{v}_k$ 分别代表第 k 次更新的 $\boldsymbol{u}_r^{(n)}$ 和 $\partial F_p^{(r,n)}(\boldsymbol{u}_r^{(n)})$，其中 $\partial F_p^{(r,n)}(\cdot)$ 为目标函数的梯度，因此得到 $\boldsymbol{u}_{k+1}$ 的更新公式

$$\boldsymbol{u}_{k+1} = \frac{\boldsymbol{u}_k - \alpha \boldsymbol{v}_k}{\sqrt{1 - 2\alpha \boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_k + \alpha^2 \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{v}_k}} \tag{8}$$

式中 α 为步长参数。调整 α 使得 $F_p^{(r,n)}(\boldsymbol{u}_{k+1}) \leqslant F_p^{(r,n)}(\boldsymbol{u}_k)$，迭代此式直至收敛；接下来化简目标函数的梯度，目标函数 $F_p^{(r,n)}(\boldsymbol{u}_k)$ 可以化简为

$$\frac{g_r^2}{2}\rho^{(n)} \left\| L^{(n)} \boldsymbol{u}_r^n \right\|_p^p - g_r \boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{y}_r^{(n)} + \frac{1}{2} g_r^2 \boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{u}_k \tag{9}$$

式中 $\boldsymbol{y}_r^{(n)} = \mathrm{vec}(\boldsymbol{Y}_r \times_1 \boldsymbol{u}_r^{(1)\mathrm{T}} \times_2 \cdots \times_{n-1} \boldsymbol{u}_r^{(n-1)\mathrm{T}} \times_{n+1} \boldsymbol{u}_r^{(n+1)\mathrm{T}} \times_{n+2} \cdots \times_N \boldsymbol{u}_r^{(N)\mathrm{T}})$，因此目标函数的梯度可表示为

$$\partial F_p^{(r,n)}(\boldsymbol{u}_k) = \begin{cases} \dfrac{g_r^2}{2}\rho^{(n)} \boldsymbol{L}^{(n)\mathrm{T}} \mathrm{SGN}(L^{(n)} \boldsymbol{u}_k) - g_r \boldsymbol{y}_r^{(n)} + g_r^2 \boldsymbol{u}_k & p = 1 \\ g_r^2 \rho^{(n)} \boldsymbol{L}^{(n)\mathrm{T}} \boldsymbol{u}_k - g_r \boldsymbol{y}_r^{(n)} + g_r^2 \boldsymbol{u}_k & p = 2 \end{cases} \tag{10}$$

式中 $\mathrm{SGN}(\boldsymbol{x}) = [\mathrm{sgn}(x_1), \mathrm{sgn}(x_2), \cdots, \mathrm{sgn}(x_J)]^{\mathrm{T}}$，其中

$$\mathrm{sgn}(x_j) = \begin{cases} 1 & x_j > 0 \\ 0 & x_j = 0 \\ -1 & x_j < 0 \end{cases} \tag{11}$$

本文算法过程示意图如图 1 所示，计算步骤如下：

(1) $H_\tau(\boldsymbol{T}) = \mathrm{fold}_{(I,\tau)} \boldsymbol{T} \times_1 \boldsymbol{S}_1 \times_2 \boldsymbol{S}_2 \cdots \times_N \boldsymbol{S}_N$，$H_\tau(\boldsymbol{W}) = \mathrm{fold}_{(I,\tau)} \boldsymbol{W} \times_1 \boldsymbol{S}_1 \times_2 \boldsymbol{S}_2 \cdots \times_N \boldsymbol{S}_N$，得到折叠后的张量 $\boldsymbol{T}$，坐标张量 $\boldsymbol{W}$，ρ，p，信号失真率 SDR 和停止阈值 υ；

(2) 算误差界限 $\varepsilon = 10^{(-\mathrm{SDR}/10)} \left\| T_\Omega \right\|_{\mathrm{F}}^2$，初始化 $\boldsymbol{X}_\Omega = T_\Omega$，将 T_Ω 的平均值赋值给 $\boldsymbol{X}_{\bar{\Omega}}$，输入平滑矩阵 $\boldsymbol{L}^{(n)}$，令初始的秩 $R = 1$；

(3) 从 1 到 N，随机初始化 $\left\{ \boldsymbol{u}_R^{(n)} \in \boldsymbol{U}^{I_n} \right\}_{n=1}^{N}$，其中 $\boldsymbol{U}^{I_n} = \left\{ \boldsymbol{u} \in \mathrm{R}^{I_n} \middle| \left\| \boldsymbol{u} \right\|_2 = 1 \right\}$ 表示因子分量的子集；

(4) 计算 CP 分解中的缩放倍数 $g_R = \left\langle \boldsymbol{X}, \boldsymbol{u}_R^{(1)} \circ \boldsymbol{u}_R^{(2)} \circ \cdots \circ \boldsymbol{u}_R^{(N)} \right\rangle$；

(5) 计算 $\boldsymbol{X}$ 与 $\boldsymbol{Z}$ 之间的误差 $E = \boldsymbol{X} - \sum_{r=1}^{R} g_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(N)}$，其中 $E_{\bar{\Omega}} = 0$；

(6) t 为迭代次数，初始值赋值为 0，最大值为 max iter，$\mu_t = \left\| \boldsymbol{E} \right\|_{\mathrm{F}}^2$ 代表 $\boldsymbol{X}$ 与 $\boldsymbol{Z}$ 之间的平方误差；

(7) t 从 0 到 max iter 重复执行步骤①～④。

①r 从 1 到 R 执行步骤 a)～e)

a) $\boldsymbol{Y}_r = E + g_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(N)}$；

b) n 从 1 到 N 依次计算：$\boldsymbol{u}_r^{(n)} = \underset{u\in U^{I_n}}{\arg\min}\, F_p^{(r,n)}(\boldsymbol{u})$；

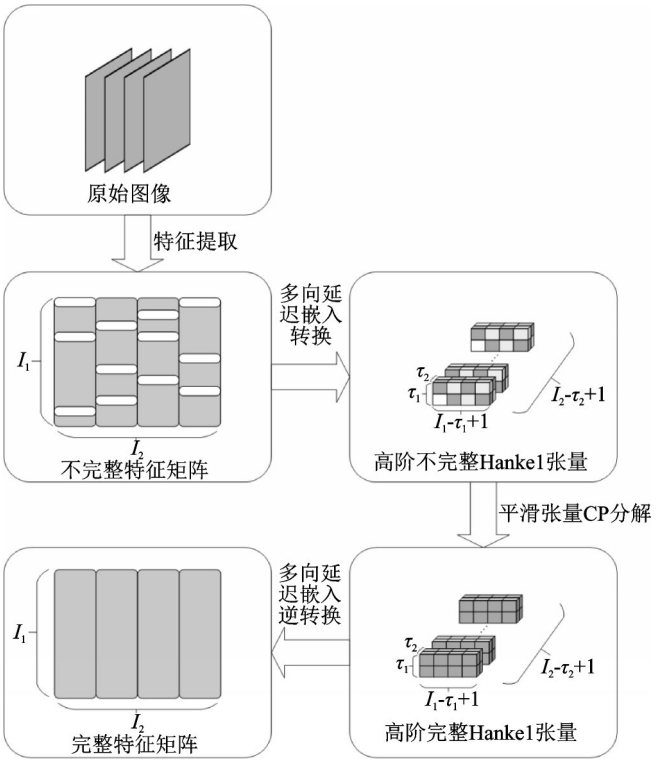

图1　基于多向延迟嵌入的平滑CP分解算法过程示意图

Fig.1　Process diagram of smooth CP decomposition algorithm based on multi-direction delay embedding

c)计算 $g_r=\dfrac{\left\langle \boldsymbol{Y}_r,\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}\right\rangle}{\left(1+\sum\limits_{n=1}^{N}\rho^{(n)}\left\|L^{(n)}\boldsymbol{u}_r^{(n)}\right\|_p^p\right)}$;

d)计算 $E=\boldsymbol{Y}_r-g_r\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}$;

e)$E_{\Omega}=0$。

②如果 $\dfrac{\left|\mu_t-\mu_{t+1}\right|}{\left|\mu_{t+1}-\varepsilon\right|}\leqslant\upsilon$,其中 $\mu_t=\left\|Z_{\Omega}^t-T_{\Omega}\right\|_{\mathrm{F}}^2$,$Z_{\Omega}^t$ 表示迭代第 t 次时的CP分解模型,依次执行以下步骤。

a)$R=R+1$;

b)随机初始化 $\left\{\boldsymbol{u}_R^{(n)}\in U^{I_n}\right\}_{n=1}^{N}$;

c)$g_R=\left\langle E,\boldsymbol{u}_R^{(1)}\circ\boldsymbol{u}_R^{(2)}\circ\cdots\circ\boldsymbol{u}_R^{(N)}\right\rangle$;

d)计算 $E=\boldsymbol{Y}_r-g_r\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}$;

e)$E_{\Omega}=0$。

③$t=t+1$;

④如果 $\mu_t\leqslant\varepsilon$,或者 $t=\max\mathrm{iter}$,跳出本层循环。

(8)$Z=\sum\limits_{r=1}^{R}g_r\boldsymbol{u}_r^{(1)}\circ\boldsymbol{u}_r^{(2)}\circ\cdots\circ\boldsymbol{u}_r^{(N)}$;

(9)$\boldsymbol{X}_{\hat{\Omega}}=\boldsymbol{Z}_{\hat{\Omega}}$;

(10)得到结果$\boldsymbol{X}$和$\boldsymbol{Z}$,经过$H_{\tau}^{-1}(\boldsymbol{X})=\text{unfold}_{(l,\tau)}(\boldsymbol{X})\times_1\boldsymbol{S}_1^{\dagger}\cdots\times_N\boldsymbol{S}_N^{\dagger}$,得到还原后的$\boldsymbol{X}$。

3 实验与分析

脑胶质瘤是一种恶性肿瘤,具有易复发、存活率低的特点[11]。世界卫生组织根据肿瘤病理的恶性程度,将胶质瘤划分成四级,其中Ⅰ~Ⅱ级为低级别胶质瘤(LGG),Ⅲ~Ⅳ级为高级别胶质瘤(HGG)[12]。中国脑胶质瘤5年病死率在全身肿瘤中仅次于胰腺癌和肺癌[13],本实验采用的是2017年MICCAI BraTS数据集[14-17],共285个病人的磁共振图像,包括75例低级别胶质瘤和210例高级别胶质瘤。每个病人的MRI影像数据包括4种模态:T1,T1ce,T2和Flair,如图2所示。

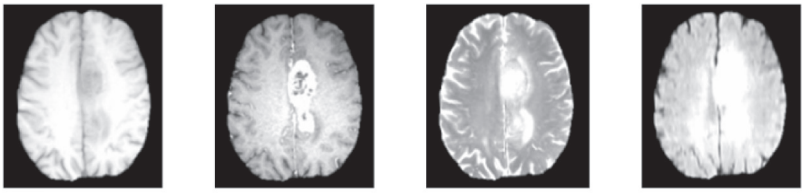

图2 BraTS2017数据集

Fig.2 BraTS2017 dataset

分别对每一种模态提取一阶统计特征和二阶纹理特征。一阶统计特征包括体积、坚固性和偏心度;二阶纹理特征包括3种全局纹理特征以及从灰度共生矩阵(Gray-level co-occurrence matrix, GLCM)、灰度游程矩阵(Gray-level run-length matrix, GLRLM)、灰度区域大小矩阵(Gray level size zone matrix, GLSZM)和邻域灰度差分矩阵(Neighborhood gray-tone difference matrix, NGTDM)中分别提取的9、13、13、5个特征,即每个模态提取46个特征,最后4个模态一共得到184个特征。

每个模态随机挑选出20%的样本,将挑选出的样本的对应模态特征做缺失处理,得到不完整的矩阵$\boldsymbol{T}$。矩阵$\boldsymbol{T}$的数据形式如图3所示,阴影部分为缺失的数据。行数据是每个病人的特征数据,每列是对应的特征,A1~A46是由模态T1提取的46个特征,B1~B46是模态T1ce提取的特征,以此类推。

首先使用数据补全算法将不完整的特征数据补全,之后使用线性核的支持向量机进行高级别胶质

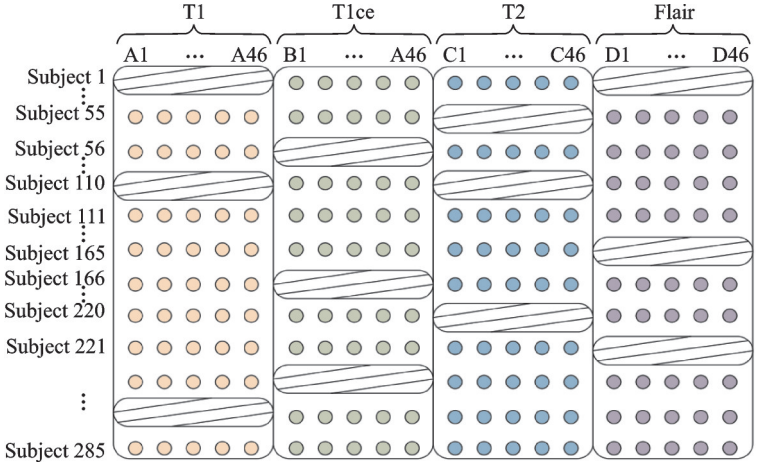

图3 特征矩阵示意图

Fig.3 Schematic diagram of characteristic matrix

瘤和低级别胶质瘤的二分类任务,最后使用平均分类准确率评估各算法性能,实验流程如图4所示。基准方法包括:补零法、邻近算法、奇异值分解算法[18]、期望最大化算法、低秩矩阵补全[19]、低秩张量补全[20]和基于多向延迟嵌入的Tucker分解算法。实验采取留出法,每次都按照4:1随机选取训练集和测试集的病人,重复进行高级别胶质瘤和低级别胶质瘤的分类实验,共重复1 000次,计算平均分类准确率,结果如表1所示。

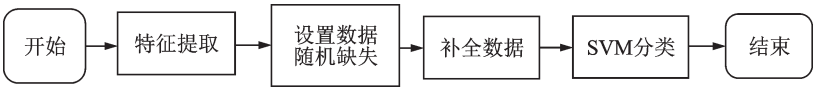

图4　实验流程图

Fig.4　Experimental flow chart

表1　实验结果

Table 1　Experimental results

实验方法	平均分类准确率/%	标准差
完整数据	91.40	3.439 1
补零法	79.77	5.171 3
邻近算法	87.81	3.603 1
期望最大化算法	89.06	3.525 8
奇异值分解算法	87.55	3.826 6
低秩补全算法	88.71	3.693 6
精确的低秩张量补全算法	82.75	4.056 2
基于多向延迟嵌入的Tucker分解算法	89.87	3.710 7
基于多向延迟嵌入的平滑张量补全算法	91.31	3.420 3

由表1可知:经过本文提出的基于多向延迟嵌入的平滑张量补全算法补全的矩阵,其分类效果相比传统算法要好,准确率达到91.31%;相比于原算法——基于多向延迟嵌入的Tucker分解算法也有一定程度上的提升,分类精度提升了1.44%。实验结果表明,改进算法减轻了由于缺失模态问题造成的特征丢失,使得分类准确率有了进一步提升。

4　结束语

本文在多向延迟嵌入Tucker分解算法的基础上,提出了一种基于多向延迟嵌入的平滑张量补全算法,主要是将数据经过多向延迟嵌入操作后进行平滑张量CP分解,最后进行多向延迟嵌入的逆向过程,得到补全的数据。实验结果表明,改进后的算法弥补了原算法Tucker分解的不足,增强了算法的降噪能力,具有更好的补齐效果和鲁棒性。算法也依然存在计算复杂度较高等不足之处,将在接下来的研究中进一步改善。

参考文献:

[1]　杜若画. 基于机器学习的缺失模态影像分类研究[D]. 青岛:山东大学, 2020.
DU Ruhua. Research on missing modal image classification based on machine learning[D]. Qingdao:Shandong University, 2020.

[2]　ZHANG Changqing, ADELI E, WU Zhengwang, et al. Infant brain development prediction with latent partial multi-view representation learning[J]. IEEE Trans Med Imaging, 2019,38(4): 909-918.

[3]　ZHANG L, ZHAO Y, ZHU Z, et al. Multi-view missing data completion[J]. IEEE Transactions on Knowledge and Data Engineering, 2018,30(7): 1.

[4]　SCHNEIDER T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of

missing values[J]. Journal of Climate, 2001, 14(5): 853-871.
[5] HASTIE T, TIBSHIRANI R, SHERLOCK G, et al. Imputing missing data for gene expression arrays[R]. USA: Statistics Department of Stanford University, 1999.
[6] 刘晨彬. 基于磁共振图像分析的神经胶质瘤分子标记物检测研究[D]. 杭州:浙江大学, 2013.
LIU Chenbin. Detection of molecular markers in glioma based on magnetic resonance image analysis[D]. Hangzhou: Zhejiang University, 2013.
[7] ASHRAPHIJUO M, WANG X, AGGARWAL V. Fundamental sampling patterns for low-rank multi-view data completion [J]. Pattern Recognition, 2020, 103: 107307.
[8] YOKOTA T, EREM B, GULER S, et al. Missing slice recovery for tensors using a low-rank model in embedded space[C]// Proceedings of the IEEE conference on computer vision and pattern recognition.[S.l.]:IEEE, 2018: 8251-8259.
[9] YOKOTA T, ZHAO Q, CICHOCKI A. Smooth PARAFAC decomposition for tensor completion[J]. IEEE Transactions on Signal Processing, 2016, 64(20): 5423-5436.
[10] KOLDA T G, BADER B W. Tensor decompositions and applications[J]. SIAM Review, 2009, 51(3): 455-500.
[11] 江涛. 脑胶质瘤精准医疗 [J]. 首都医科大学学报, 2020, 41(5): 854-859.
JIANG Tao. Precise medical treatment of glioma [J]. Journal of Capital Medical University, 2020, 41(5): 854-859.
[12] LOUIS D N, OHGAKI H, WIESTLER O D, et al. The 2007 WHO classification of tumours of the central nervous system [J]. Acta Neuropathologica, 2007, 114(2): 97-109.
[13] 吴强, 何泽鲲, 刘琚, 等. 基于机器学习的脑胶质瘤多模态影像分析[J]. 山东大学学报 (医学版), 2020, 58(8): 81-87.
WU Qiang, HE Zekun, LIU Ju, et al. Multimodal image analysis of glioma based on machine learning[J]. Journal of Shandong University (Medical Edition), 2020, 58(8): 81-87.
[14] BAKAS S, AKBARI H, SOTIRAS A, et al. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection[J]. The Cancer Imaging Archive, 2017.DOI: http://10.7937/K9/TCIA.2017.KLXWJJ1Q.
[15] MENZE B H, JAKAB A, BAUER S, et al. The multimodal brain tumor image segmentation benchmark (BRATS)[J]. IEEE Transactions on Medical Imaging, 2014, 34(10): 1993-2024.
[16] BAKAS S, AKBARI H, SOTIRAS A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features[J]. Scientific Data, 2017, 4: 170117.
[17] BAKAS S, REYES M, JAKAB A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression ASSESSMENT , and overall survival prediction in the BRATS challenge[EB/OL]. (2018-11-05)[2020-11-15]. http://export.arxiv.org/abs/1811.02629.
[18] GOLUB G H, REINSCH C. Singular value decomposition and least squares solutions[M]. Berlin, Heidelberg: Springer, 1971: 134-151.
[19] CANDES E J, RECHT B. Exact low-rank matrix completion via convex optimization[C]//Proceedings of 2008 46th Annual Allerton Conference on Communication, Control, and Computing. [S.l.]: IEEE, 2008: 806-812.
[20] LIU J, MUSIALSKI P, WONKA P, et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 208-220.

作者简介:

刘琚(1965-),男,教授,博士生导师,研究方向:无线通信中空时信号处理技术、多媒体通信与网络传输技术,E-mail:juliu@sdu.edu.cn。

杜若画(1995-),**通信作者**,女,硕士研究生,研究方向:医学图像处理,E-mail:466079344@qq.com。

吴强(1979-),男,副教授,硕士研究生导师,研究方向:机器学习、生物医学信号处理。

何泽鲲(1996-),男,硕士研究生,研究方向:医学图像处理。

于璐跃(1997-),女,博士研究生,研究方向:医学图像处理。