文章编号：2095-4980(2022)09-0965-08

链接文档中基于子空间分解的高效谱聚类算法

原　虹，赵　丽，王溢琴

(晋中学院 信息技术与工程系，山西 晋中 030619)

摘　要：提出了一种基于子空间分解的高效谱聚类算法。首先，基于共识信息和特定域信息的矩阵分解将链接文档划分为3个子空间，然后对子空间添加正则化项建模共识信息和特定域信息对聚类的不同影响，并采用交替优化方法实现谱聚类。考虑到谱聚类的复杂性，提出了一种带曲线搜索的梯度下降法加速求解过程。3个真实数据集上的实验结果表明，所提算法在聚类质量和效率方面始终明显优于目前典型的基线算法，且对输入参数不敏感。

关键词：链接文档；子空间分解；谱聚类；梯度下降法；基线算法

中图分类号：TP393　　**文献标志码：**A　　**doi：**10.11805/TKYDA2020183

Efficient spectral clustering algorithm based on subspace decomposition in linked documents

YUAN Hong，ZHAO Li，WANG Yiqin

(School of Information Technology and Engineering，Jinzhong University，Jinzhong Shanxi 030619，China)

Abstract：An efficient spectral clustering algorithm based on subspace decomposition is proposed. Firstly, based on the matrix decomposition of consensus information and domain specific information, the linked documents are divided into three subspaces, the different effects of consensus information and domain specific information on clustering are modeled by adding regularization items to the subspaces, and the alternative optimization method is utilized to achieve spectral clustering. In addition, considering the complexity of spectral clustering, a gradient descent method with curvilinear search is proposed to accelerate the solution process. Experimental results on three real datasets show that the proposed algorithm is superior to the current typical baseline algorithm in terms of clustering quality and efficiency, and is insensitive to input parameters.

Keywords：linked documents；subspace decomposition；spectral clustering；gradient descent method；baseline algorithm

文本数据和链接数据是指具有丰富文本内容的互连对象的集合，也称为链接文档。对链接文档进行聚类在很多现实场景中都有应用[1]，如，可以在搜索引擎中使用聚类来促进文档组织，改进搜索结果排名，以及更快地进行信息检索。聚类还可以在社交网站(Facebook、Twitter等)中用于以群组为目标的广告和营销[2]。

人们已经对用于链接文档的聚类方法进行了广泛研究，其中，代表性的聚类方法主要有：概率方法[3-5]、矩阵分解方法[6-7]和基于相似性和多视图的方法[8-10]。但以上的所有方法在聚类过程中主要关注共识信息的建模，没有考虑文档间的信息差异，仅将区别于共识信息的特定域信息视作误差。为此，本文提出一种改进的聚类方法[11]，将任意域的数据看作三部分：共识域、特定域和误差，并通过对共识信息和特定域信息的矩阵分解进行聚类建模。具体地，将链接数据自然看作一个图，而文本数据则通过词语相似度关系来构建图，采用图的对称矩阵分解将其顶点映射至低维子空间中；然后基于共识信息和特定域信息的矩阵分解产生3个子空间：文本和链接数据之间的共识子空间、文本特定域的子空间和链接数据特定域的子空间；最后，对子空间分解结果进行正则化处理，凸显特定域信息对于聚类建模的影响，从而得到一个优化目标为共识信息最大化的聚类模型。

收稿日期：2020-04-30；**修回日期：**2020-10-05

基金项目：新工科背景下“双师”型师资队伍建设2019年第一批产学合作协同育人资助项目(201901040022)；山西省自然科学基金资助项目(201901D22111)；山西省教育科学“十三五”规划2020年度‘互联网+教育’专项课题(HLW-20111)。

为求解所提模型的优化问题，在子空间上施加正交性约束将其转化为谱聚类[11]。然后通过标准的谱聚类子程序来解决该优化问题。考虑到谱聚类的计算过程依赖于稠密图的特征分解，开销较大，为此，提出利用数据稀疏性和子空间的低维特点来部署一个带曲线搜索的梯度下降方法。3个真实数据集上的仿真实验结果也表明了本文算法的有效性。

1 问题描述

定义由文本数据和链接数据组成的n个链接文档集合为：$C=(\boldsymbol{X},\boldsymbol{A})$。其中，$\boldsymbol{X}$表示链接数据，$\boldsymbol{A}$表示文本数据。$\boldsymbol{X}$为$n\times n$矩阵，其中的每个元素$X_{ij}$表示连接第$i$个文档和第$j$个文档的链接数量。$\boldsymbol{A}$为$n\times m$矩阵，每行表示一个文档，每列表示大小为$m$的字典[12]中的一个单词或字的特征，$A_{ij}$表示在第$i$个文档中第$j$个字的出现频率。给定任意的链接文档集$C$和聚类数目$k$，本文研究的聚类问题是：如何将$C$划分为$k$个不相交的聚类，使文本数据和链接数据之间的共识信息最大化。

2 本文方法

本文提出的改进聚类方法首先在图上引入对称矩阵分解，构成聚类模型的基础；然后，基于共识信息和特定域信息对文本和链接数据执行分解；最后，采用正则化方法锐化子空间，给出模型的最终目标函数。

2.1 对称矩阵分解

给定一个图的邻接矩阵$\boldsymbol{W}$，它的标准化邻接矩阵和标准化拉普拉斯算子可分别定义为：$\tilde{\boldsymbol{W}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$和$\boldsymbol{L}=\boldsymbol{I}-\tilde{\boldsymbol{W}}$。其中，$\boldsymbol{I}$为单位矩阵，是全一列向量，$\boldsymbol{D}=diag(\boldsymbol{W}1)$。则在正交性约束条件下图的对称矩阵分解问题可建模为如下的优化问题：

$$\min_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}\|\tilde{\boldsymbol{W}}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_{\mathrm{F}}^{2} \tag{1}$$

式中：$\|\cdot\|_{\mathrm{F}}$为Frobenius范数[13]；$\boldsymbol{U}$的行可看作是图顶点在潜在的低维子空间中的投影，这些投影(即低维表示)编码了关于顶点聚类的判别信息，可被反馈至K-means算法中找到最终的聚类结果。由于正交性约束，问题(1)等价于求解给定图的谱聚类问题，即$\min\limits_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}tr(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{U})$。

2.2 基于共识信息和特定域信息的分解

在文本数据上构造一个余弦相似图$\boldsymbol{Y}$(在链接数据上做类似处理，可得余弦相似图$\boldsymbol{X}$)，其中的任意元素$\boldsymbol{Y}_{ij}=\dfrac{\boldsymbol{A}_{i\cdot}\boldsymbol{A}_{j\cdot}^{\mathrm{T}}}{\|\boldsymbol{A}_{i\cdot}\|_2\|\boldsymbol{A}_{j\cdot}\|_2}$，$\boldsymbol{A}_{i\cdot}$表示$\boldsymbol{A}$的第$i$行。进一步地，分别定义$\tilde{\boldsymbol{X}}$和$\tilde{\boldsymbol{Y}}$为$\boldsymbol{X}$和$\boldsymbol{Y}$的标准化邻接矩阵。标准化文本数据$\tilde{\boldsymbol{A}}$定义为：$\boldsymbol{A}_{i\cdot}=(\boldsymbol{Y}_{i\cdot}1)^{-1/2}\dfrac{\boldsymbol{A}_{i\cdot}}{\|\boldsymbol{A}_{i\cdot}\|_2}$，有$\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\mathrm{T}}$。然后对链接数据和文本数据执行如下的对称矩阵分解：

$$\tilde{\boldsymbol{X}}=\boldsymbol{U}_1\boldsymbol{U}_1^{T}+\boldsymbol{E}_1,\tilde{\boldsymbol{Y}}=\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}}+\boldsymbol{E}_2 \tag{2}$$

式中$\boldsymbol{E}_i$表示链接数据或文本数据的重构误差矩阵，文中基于$\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}$约束条件下的Frobenius范数对其进行最小化。为了描述文本数据和链接数据中的共识信息和特定域信息，将子空间分解为一个共识子空间$\boldsymbol{U}\in\mathbf{R}^{n\times k}$和两个特定域子空间$\boldsymbol{U}_i\in\mathbf{R}^{n\times k_i}$：子空间$\leftarrow[\boldsymbol{U},\boldsymbol{U}_i]$

对于文本数据和链接数据，可以进行如下的共识信息和特定域信息的分解：

$$\begin{cases}\tilde{\boldsymbol{X}}=\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}+\boldsymbol{E}_1\\ \tilde{\boldsymbol{Y}}=\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}}+\boldsymbol{E}_2\end{cases} \tag{3}$$

式中：$\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$表示文本数据和链接数据之间的共识信息；$\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i(i=1,2)$表示文本数据和链接数据中的特定域信息。在式(3)中，$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$和$\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}(i=1,2)$。为了从$\boldsymbol{U}$中使用共识信息实现聚类，需解决如下的优化问题：

$$\begin{cases}\min\|\boldsymbol{E}_1\|_{\mathrm{F}}^{2}+\|\boldsymbol{E}_2\|_{\mathrm{F}}^{2}\\ \text{满足条件:}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}\text{和}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}(i=1,2)\end{cases} \tag{4}$$

2.3 基于正则化的聚类建模

根据共识信息和特定域信息的定义，3 个子空间 $\boldsymbol{U}$ 和 $\boldsymbol{U}_i(i=1,2)$ 之间应该彼此不同。因此，为了真实建模不同子空间的差异对聚类的影响以及避免不同子空间之间的信息混杂所导致的噪声，添加正则化项[8]至式(4)中的目标函数：

$$d(\boldsymbol{U}_1,\boldsymbol{U}_2)=\left\|\frac{K_{\boldsymbol{U}_1}}{\|K_{\boldsymbol{U}_1}\|_{\mathrm{F}}^2}-\frac{K_{\boldsymbol{U}_2}}{\|K_{\boldsymbol{U}_2}\|_{\mathrm{F}}^2}\right\|_{\mathrm{F}}^2 \tag{5}$$

式中 $K_{\boldsymbol{U}_i}=\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}$ 为线性相似核。与文献[8]相反，本文要求 $d(\boldsymbol{U}_1,\boldsymbol{U}_2)$ 尽可能大。因此还需对 $(\boldsymbol{U},\boldsymbol{U}_1)$ 和 $(\boldsymbol{U},\boldsymbol{U}_2)$ 添加类似的正则化项，最终得到本文的聚类模型为：

$$\begin{cases}\min\sum_{i=1}^{2}\left[\|\boldsymbol{E}_i\|_{\mathrm{F}}^2-\lambda d(\boldsymbol{U},\boldsymbol{U}_i)\right]-\lambda d(\boldsymbol{U}_1,\boldsymbol{U}_2)\\ \text{满足条件:}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}\text{和}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}(i=1,2)\end{cases} \tag{6}$$

式中 λ 为正则化系数，取值大于 0。

3 优化方法

通过优化式(6)中的目标函数进行模型求解。首先对目标函数进行简化：

$$\Gamma(\boldsymbol{U},\boldsymbol{U}_1,\boldsymbol{U}_2)=\sum_{i=1}^{2}\left[\|\boldsymbol{E}_i\|_{\mathrm{F}}^2-\lambda d(\boldsymbol{U},\boldsymbol{U}_i)\right]-\lambda d(\boldsymbol{U}_1,\boldsymbol{U}_2) \tag{7}$$

将 $\|K_{\boldsymbol{U}}\|_{\mathrm{F}}^2=k$ 和 $\|K_{\boldsymbol{U}_i}\|_{\mathrm{F}}^2=k_i(i=1,2)$ 代入式(6)中，忽略加法和依赖于 k 和 $k_i(i=1,2)$ 的定标常数，可得：

$$d(\boldsymbol{U}_1,\boldsymbol{U}_2)=-tr(\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}}) \tag{8}$$

类似地，有：

$$\begin{cases}\|\boldsymbol{E}_1\|_{\mathrm{F}}^2=-tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U})-tr(\boldsymbol{U}_1^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)-d(\boldsymbol{U},\boldsymbol{U}_1)\\ \|\boldsymbol{E}_2\|_{\mathrm{F}}^2=-tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{Y}}\boldsymbol{U})-tr(\boldsymbol{U}_2^{\mathrm{T}}\tilde{\boldsymbol{Y}}\boldsymbol{U}_2)-d(\boldsymbol{U},\boldsymbol{U}_2)\end{cases} \tag{9}$$

最后，再次使用相同的正则化系数 λ，则目标函数变成：

$$J(\boldsymbol{U},\boldsymbol{U}_1,\boldsymbol{U}_2)=tr[\boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}})\boldsymbol{U}]+tr(\boldsymbol{U}_1^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)+tr(\boldsymbol{U}_2^{\mathrm{T}}\tilde{\boldsymbol{Y}}\boldsymbol{U}_2)-\lambda\left\{tr[\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}+\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})]+tr(\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})\right\} \tag{10}$$

此时，本文要优化的聚类模型转化为最大化式(10)中的目标函数：

$$\begin{cases}\max J(\boldsymbol{U},\boldsymbol{U}_1,\boldsymbol{U}_2)\\ \text{满足条件:}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}\text{和}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}(i=1,2)\end{cases} \tag{11}$$

将式(11)转换为谱聚类进行求解，然后提出一种交替优化方法来取代代价昂贵的特征分解，以提高计算效率。

3.1 迭代过程

正交约束条件下式(10)所示的目标函数可通过交替优化来实现最大化。即在优化 $\boldsymbol{U}$ 的过程中固定其他参数不变，直到优化过程收敛，优化结束；然后采用同样的思路优化 $\boldsymbol{U}_i(i=1,2)$；最后，在 $\boldsymbol{U}$ 上执行 K-means 算法，将聚类数目设置为 $\boldsymbol{U}$ 中的列数(即 k)，从而得到文档的聚类结果。具体迭代过程如下。

3.1.1 固定 $\boldsymbol{U}_1$ 和 $\boldsymbol{U}_2$，优化 $\boldsymbol{U}$

对 $\boldsymbol{U}$ 进行优化的目标函数为：

$$\boldsymbol{F}(\boldsymbol{U})=tr[\boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}})\boldsymbol{U}]-\lambda tr(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\sum_{i=1}^{2}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})=tr\left[\boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}}-\lambda\sum_{i=1}^{2}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})\boldsymbol{U}\right] \tag{12}$$

式中 $\boldsymbol{U}=\arg\max_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}\boldsymbol{F}(\boldsymbol{U})$ 可以通过求解和图 $(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}}-\lambda\sum_{i=1}^{2}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})$ 上的标准谱聚类来获得。它由和图上的 k 个最大特征向

量组成。$(\tilde{\boldsymbol{X}}-\lambda\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}})$和$(\tilde{\boldsymbol{Y}}-\lambda\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})$可看作在过滤了特定域信息后，仍存在于文本数据和链接数据中的共识信息，而和图则可以看作是这两种共识信息的叠加。

3.1.2 固定$\boldsymbol{U}$和$\boldsymbol{U}_2$，优化$\boldsymbol{U}_1$

对$\boldsymbol{U}_1$进行优化的目标函数是：

$$\boldsymbol{F}_1(\boldsymbol{U}_1)=tr(\boldsymbol{U}_1^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)-\lambda tr\left[\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})\right]=tr\left\{\boldsymbol{U}_1^{\mathrm{T}}\left[\tilde{\boldsymbol{X}}-\lambda(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})\right]\boldsymbol{U}_1\right\} \tag{13}$$

式中$\boldsymbol{U}_1=\arg\max_{\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1=\boldsymbol{I}}\boldsymbol{F}_1(\boldsymbol{U}_1)$由求和图$[\tilde{\boldsymbol{X}}-\lambda(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}})]$的$k_1$个最大特征向量组成。和图可看作在链接数据中过滤了共识信息后的特定域信息。而进一步减去$\lambda\boldsymbol{U}_2\boldsymbol{U}_2^{\mathrm{T}}$是为了使链接数据的特定域部分与文本数据的特定域部分区别开来。

3.1.3 固定$\boldsymbol{U}$和$\boldsymbol{U}_1$，优化$\boldsymbol{U}_2$

对$\boldsymbol{U}_2$进行优化的目标函数是：

$$\boldsymbol{F}_2(\boldsymbol{U}_2)=tr(\boldsymbol{U}_2^{\mathrm{T}}\tilde{\boldsymbol{Y}}\boldsymbol{U}_2)-\lambda tr\left[\boldsymbol{U}_2^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}})\boldsymbol{U}_2\right]=tr\left\{\boldsymbol{U}_2^{\mathrm{T}}\left[\tilde{\boldsymbol{Y}}-\lambda(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}})\right]\boldsymbol{U}_2\right\} \tag{14}$$

式中$\boldsymbol{U}_2=\arg\max_{\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2=\boldsymbol{I}}\boldsymbol{F}_2(\boldsymbol{U}_2)$是由和图$[\tilde{\boldsymbol{Y}}-\lambda(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}})]$的$k_2$个最大特征向量组成。从上述的迭代过程中可以看出，与忽略特定域信息的现有方法[5−6,8−9]相比，本文方法可以明确区分文本特定域信息和链接数据特定域信息对聚类建模的影响，因此可以保证文本和链接数据之间的共识部分更加精准可靠，聚类结果更准确。

3.2 初始化

迭代过程需要初始化。文中给出一种确定性的初始化方案。具体而言，先将$\boldsymbol{U}$初始化为：

$$\boldsymbol{U}=\arg\max_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}\tilde{\boldsymbol{F}}(\boldsymbol{U})=\arg\max_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}tr[\boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}})\boldsymbol{U}] \tag{15}$$

对比3.1.1节中$\boldsymbol{U}$的优化过程可知，式(15)中的$(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{Y}})$混合了文本和链接数据的共识信息和特定域信息，即初始化后的$\boldsymbol{U}$可以看作共识信息的未净化版本。类似地，$\boldsymbol{U}_1$和$\boldsymbol{U}_2$可以初始化为：

$$\boldsymbol{U}_i=\arg\max_{\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{I}}tr\left[\boldsymbol{U}_i^{\mathrm{T}}(\tilde{\boldsymbol{X}}-\lambda\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\boldsymbol{U}_i\right] \tag{16}$$

3.3 带曲线搜索的梯度下降

式(11)所示的优化问题可用谱聚类来表示并使用其标准子程序进行求解，但由于式(13)~(16)中的和图都是稠密的，计算量很大，因此本文提出一种带曲线搜索(Curvilinear search)的梯度下降方法进行求解。该方法能充分利用求和图的低维结构和数据稀疏性有效解出目标函数及其导数，并能保证沿负梯度下降的每一步都满足正交性约束。以式(12)中的优化子问题为例说明该方法的主要过程。在梯度下降过程的每次迭代中，对于任意给定的一个可行点$\boldsymbol{U}$(即$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$)，它的负目标函数$-F(\boldsymbol{U})$及其梯度$\boldsymbol{G}$可计算为：

$$\begin{cases}-\boldsymbol{F}(\boldsymbol{U})=-tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U})-\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^2+\lambda\sum_{i=1}^{2}\|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^2\\ \boldsymbol{G}=\nabla[-\boldsymbol{F}(\boldsymbol{U})]=-(\boldsymbol{X}+\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\mathrm{T}}-\lambda\sum_{i=1}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})\boldsymbol{U}=-\tilde{\boldsymbol{X}}\boldsymbol{U}-\tilde{\boldsymbol{A}}(\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U})+\lambda\sum_{i=1}^{2}\boldsymbol{U}_i(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U})\end{cases} \tag{17}$$

与全矩阵$(\tilde{\boldsymbol{X}}+\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{T}-\lambda\sum_{i=1}^{2}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})$进行特征分解的计算开销相比，由于$\tilde{\boldsymbol{X}}$和$\tilde{\boldsymbol{A}}$是稀疏的，$\boldsymbol{U}$和$\boldsymbol{U}_i(i=1,2)$是低维的。因此，采用如上的方法计算$-\boldsymbol{F}(\boldsymbol{U})$和$\boldsymbol{G}$是非常有效的。

接下来，构造一个斜对称矩阵$\boldsymbol{F}=\boldsymbol{G}\boldsymbol{U}^{\mathrm{T}}-\boldsymbol{U}\boldsymbol{G}^{\mathrm{T}}$(即有$\boldsymbol{F}^{\mathrm{T}}=-\boldsymbol{F}$)，并在Stiefel流形[10]$M=\{\boldsymbol{U}:\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{U}\}$上沿光滑曲线$Q(\tau)=\left(\boldsymbol{I}+\frac{\tau}{2}\boldsymbol{F}\right)^{-1}\left(\boldsymbol{I}-\frac{\tau}{2}\boldsymbol{F}\right)\boldsymbol{U}$搜索下一个新的可行点。因为$\frac{\mathrm{d}Q(\tau)}{\mathrm{d}\tau}|_{\tau=0}$等于$-\boldsymbol{G}$的投影在当前点$\boldsymbol{U}=Q(0)$处的切线空间$M$上，曲线$Q(\tau)(\tau\geqslant 0)$是当前点附近的下降路径。因此，要找到下一个新的可行点$\boldsymbol{U}$(当$-\boldsymbol{F}(\boldsymbol{U})$减小时)，只需沿曲线找到适当的步长大小$\tau>0$。为此，选择Barzilai Borwein(BB)步长[14]来加速非单调线性搜索。特别地，第t次迭代的BB步长被设置为：

$$\tau^{(t)}=\frac{tr[(\boldsymbol{U}^{(t)}-\boldsymbol{U}^{(t-1)})^{\mathrm{T}}(\boldsymbol{U}^{(t)}-\boldsymbol{U}^{(t-1)})]}{\left|tr\left[(\boldsymbol{U}^{(t)}-\boldsymbol{U}^{(t-1)})^{\mathrm{T}}(\tilde{\boldsymbol{G}}^{(t)}-\tilde{\boldsymbol{G}}^{(t-1)})\right]\right|}\delta^{h} \tag{18}$$

式中 h 等于满足以下条件的最小整数：

$$-\boldsymbol{F}\left[Q^{(t)}(\tau^{(t)})\right]\leqslant r^{(t)}+\rho\tau^{(t)}\frac{\mathrm{d}}{\mathrm{d}r}\left\{-\boldsymbol{F}[Q^{(t)}(\tau)]\right\}|_{\tau=0} \tag{19}$$

式中：ρ,η,δ 为正常数；$s^{(0)}=1$，$s^{(t)}=\eta s^{(t-1)}+1$；$r^{(0)}=-F(\boldsymbol{U}^{(0)})$，$r^{(t)}=\left[\eta s^{(t-1)}r^{(t-1)}-\boldsymbol{F}(\boldsymbol{U}^{(t)})\right]/s^{(t)}$；$Q^{(t)}(\tau)=\left(\boldsymbol{I}+\frac{\tau}{2}\boldsymbol{F}^{(t)}\right)^{-1}\left(\boldsymbol{I}-\frac{\tau}{2}\boldsymbol{F}^{(t)}\right)\boldsymbol{U}^{(t)}$，$\tilde{\boldsymbol{G}}=\boldsymbol{G}-\boldsymbol{U}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{U}$，$\frac{\mathrm{d}}{\mathrm{d}\tau}\left\{-\boldsymbol{F}\left[Q(\tau)\right]\right\}|_{\tau=0}=-\frac{1}{2}\left\|\tilde{\boldsymbol{G}}\right\|_{\mathrm{F}}^{2}$。

为了保证完备性，将算法 1 中的曲线搜索过程作为求解模型的子程序。类似地，式(13)~(16)中的优化子问题也可以用算法 1 求解。本文提出的用于文本数据和链接数据的聚类算法在算法 2 中给出，缩写为 SACTL(Spectral Algorithm for Clustering Text and Link data)。为了提高效率，文中直接在 $\tilde{\boldsymbol{A}}$ 上操作而不是显式地计算 $\boldsymbol{Y}$，算法 2 中涉及的目标函数可有效地计算如下：

$$\begin{cases}\tilde{\boldsymbol{F}}=tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U})+\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^{2},\\ \tilde{\boldsymbol{F}}_1=tr(\boldsymbol{U}_1{}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)-\lambda\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_1\|_{\mathrm{F}}^{2}\\ \tilde{\boldsymbol{F}}_2=\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}_2\|_{F}^{2}-\lambda\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_2\|_{F}^{2}\\ \boldsymbol{F}=tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U})+\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^{2}-\lambda\sum_{i=1}^{2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_i\|_{\mathrm{F}}^{2}\\ \boldsymbol{F}_1=tr(\boldsymbol{U}_1{}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)-\lambda(\|\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^{2}+\|\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_2\|_{\mathrm{F}}^{2})\\ \boldsymbol{F}_2=\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}_2\|_{\mathrm{F}}^{2}-\lambda(\|\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^{2}+\|\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_1\|_{\mathrm{F}}^{2})\\ J(\boldsymbol{U},\boldsymbol{U}_1,\boldsymbol{U}_2)=tr(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U})+tr(\boldsymbol{U}_1{}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{U}_1)+\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}\|_{\mathrm{F}}^{2}+\|\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}_2\|_{\mathrm{F}}^{2}-\lambda(\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_1\|_{\mathrm{F}}^{2}+\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_2\|_{\mathrm{F}}^{2}+\|\boldsymbol{U}_1{}^{\mathrm{T}}\boldsymbol{U}_2\|_{\mathrm{F}}^{2})\end{cases} \tag{20}$$

Algorithm 1 Curvilinear search with BB steps

Input: $\boldsymbol{U}^{(0)};\tau,\rho,\eta,\delta,\varepsilon\in(0,1)$

Output: $\boldsymbol{U}^{(t)}$

1: Initialize $s^{(0)}=1$, $r^{(0)}=-F(\boldsymbol{U}^{(0)})$, $t=0$

2: while $\|\tilde{\boldsymbol{G}}^{(t)}\|_{\mathrm{F}}>\varepsilon$ do

3:　while $-F|\boldsymbol{Q}^{(t)}(\tau)|>r^{(t)}-\frac{\rho\tau}{2}\left\|\tilde{\boldsymbol{G}}^{(t)}\right\|_{\mathrm{F}}^{2}$ do

4:　　$\tau\leftarrow\delta\tau$

5:　$\boldsymbol{U}^{(t+1)}=\boldsymbol{Q}^{(t)}(\tau)$

6:　$s^{(t+1)}=\eta s^{(t)}+1$

7:　$r^{(t+1)}=[\eta s^{(t)}r^{(t)}-F(\boldsymbol{U}^{(t+1)})]/s^{(t+1)}$

8:　$\xi=\frac{tr\left[(\boldsymbol{U}^{(t+1)}-\boldsymbol{U}^{(t)})^{\mathrm{T}}(\boldsymbol{U}^{(t+1)}-\boldsymbol{U}^{(t)})\right]}{\left|tr\left[(\boldsymbol{U}^{(t+1)}-\boldsymbol{U}^{(t)})^{\mathrm{T}}(\tilde{\boldsymbol{G}}^{(t+1)}-\tilde{\boldsymbol{G}}^{(t)})\right]\right|}$

9:　$\tau\leftarrow\max\{\min(\xi,\tau_{\max}),\tau_{\min}\}$

10:　$t\leftarrow t+1$

Algorithm 2 SACTL

Input: $\tilde{\boldsymbol{X}},\tilde{\boldsymbol{A}},\lambda,k$ and $k_i,i=1,2$

Output: $\boldsymbol{U}$

1: Find $\boldsymbol{U}=\arg\min_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}-\tilde{\boldsymbol{F}}(\boldsymbol{U})$ by Algorithm 1

2: Fix $\boldsymbol{U}$ and find $\boldsymbol{U}_1=\arg\min_{\boldsymbol{U}_1{}^{\mathrm{T}}\boldsymbol{U}_1=\boldsymbol{I}}-\tilde{\boldsymbol{F}}_1(\boldsymbol{U}_1)$ by Algorithm 1

3: Fix $\boldsymbol{U}$ and find $\boldsymbol{U}_2=\arg\min_{\boldsymbol{U}_2{}^{\mathrm{T}}\boldsymbol{U}_2=\boldsymbol{I}}-\tilde{\boldsymbol{F}}_2(\boldsymbol{U}_2)$ by Algorithm 1

4: repeat

5:　Fix $\boldsymbol{U}_i$, $i=1,2$, and find $\boldsymbol{U}=\arg\min_{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}}-\boldsymbol{F}(\boldsymbol{U})$ by Algorithm 1

6:　Fix $\boldsymbol{U}$, $\boldsymbol{U}_2$, and find $\boldsymbol{U}_1=\arg\min_{\boldsymbol{U}_1{}^{\mathrm{T}}\boldsymbol{U}_1=\boldsymbol{I}}-\boldsymbol{F}_1(\boldsymbol{U}_1)$ by Algorithm 1

7:　Fix $\boldsymbol{U}$, $\boldsymbol{U}_1$, and find $\boldsymbol{U}_2=\arg\min_{\boldsymbol{U}_2{}^{\mathrm{T}}\boldsymbol{U}_2=\boldsymbol{I}}-\boldsymbol{F}_2(\boldsymbol{U}_2)$ by Algorithm 1

8: until $J(\boldsymbol{U},\boldsymbol{U}_1,\boldsymbol{U}_2)$ converges

4　仿真实验

4.1　数据集

选择目前最为典型的 3 个文档数据集[5]进行仿真实验：Cora,Citeseer 和 PubMed。对于文本数据 Cora 和 Citeseer，采用矩阵 $\boldsymbol{A}$ 中元素的值为 1 或 0 表示字典中的单词在文档中至少出现一次，而对于链接数据 PubMed，则采用矩阵 $\boldsymbol{A}$ 中元素的值进行单词计数。表 1 为 3 种数据集的主要统计数据。其中，字特征数是指字典的大小，总字数是指每个文档中使用的字特征数之和，即 $nnz(\boldsymbol{A})$。

表 1　3 种文档数据集概况

Table1 Overview of three document datasets

dataset	Cora	Citeseer	PubMed
clusters	7	6	3
documents	2 708	3 312	19 717
links	5 429	4 608	44 335
word features	1 433	3 703	4 209
total words	49 216	105 165	1 333 397

4.2 评价指标

使用6种评价指标对比SACTL算法和其他算法在聚类质量方面的性能：精确性、召回率、F值、平均熵、归一化互信息(Normalized Mutual Information，NMI)和ARI(Adjusted Rand Index)[14]。

4.3 基线算法

使用目前常见的6种基线算法：PMTLM-DC[5],PCL-DC[10],RRMF[6],RMVSC[9],PMVSC,CMVSC[8]和本文提出的SACTL算法进行了性能比较。其中，对于文献[8-9]中的基线算法，使用余弦核 $k(x,y)=\frac{x^{\mathrm{T}}y}{\|x\|_2\|y\|_2}$ 构造文本数据所需的相似度矩阵。而对于其他所有的基线算法，则直接遵循原作者的思想。将所有算法在Matlab 2018b平台上编程实现，并在配置为Linux OS,Intel Xeon 2.80 GHz CPU和25 GB RAM的机器上进行测试。

4.4 实验结果

表2 不同数据集上的聚类数量

Table 2 Number of clusters on different datasets

dataset	Cora	Citeseer	PubMed
k	7	6	3
k_1	3	3	1
k_2	14	12	6

4.4.1 实验设置

整个实验中聚类数量k设置为一个真实值。为文本数据选择的聚类数量k_2比链接数据选择的聚类数量k_1更大，因为文本数据通常比链接数据更复杂。特别地，k_1设置为k的大约一半，而k_2在所有实验中设置为k的2倍。详细情况见表2。在所有实验中，正则化系数λ设为固定值0.7。在曲线搜索中采用文献[14]中的默认值作为步长。将算法的最大迭代次数设置为5，收敛阈值设置为10^{-5}，并报告10次K-means的平均结果。以所有算法的“均值±标准”的形式报告平均质量结果。

4.4.2 聚类质量分析

图1~3为不同算法在3个数据集上的性能比较结果。本文无法在最大的数据集PubMed上获得RMVSC方法的结果，因为它耗尽了内存。如图1~3所示，本文方法始终明显优于所有基线算法。与最佳的基线算法结果相比，在Cora数据集上，SACTL算法的F值、NMI和ARI分别相对提高了15.2%,19.5%和23.2%；在Citeseer数据集上，SACTL算法的F值、NMI和ARI分别提高了4.8%,6.4%和7.5%；在PubMed数据集中，SACTL算法的F值、NMI和ARI分别提高了6.2%,5.1%和17.7%。值得注意的是，所有现有基线都忽略了文本和链接中特定域信息的区别，这表明有必要对这些信息进行建模，以便更好地捕捉共识。

此外可以看到，与PMVSC/CMVSC算法相比，SACTL算法在Cora,Citeseer和PubMed上所获得的NMI聚类质量比PMTLM-DC分别提高29.9%,39.1%和56.6%，如图3(a)所示；与RMVSC算法相比，SACTL算法在Cora和Citeseer上的ARI聚类性能分别提高了305.1%和114.8%，如图3(b)所示。

5 结论

从多个数据域寻求共识信息的想法在社区网络中得到了很好的认可和广泛运用。但特定域信息的重要作用通常被忽略。本文的工作表明，通过在文本数据和链接数据的上下文中明确地建模特定域的区别，可以更好地分离共识部分并实现改进聚类的性能。仿真实验结果表明，所提的SACTL算法的聚类质量和效率都要优于现有最佳的基线算法。下一步工作中将研究面向缺失值处理的子空间聚类算法，进一步提升聚类的质量和拓展聚类的应用价值。

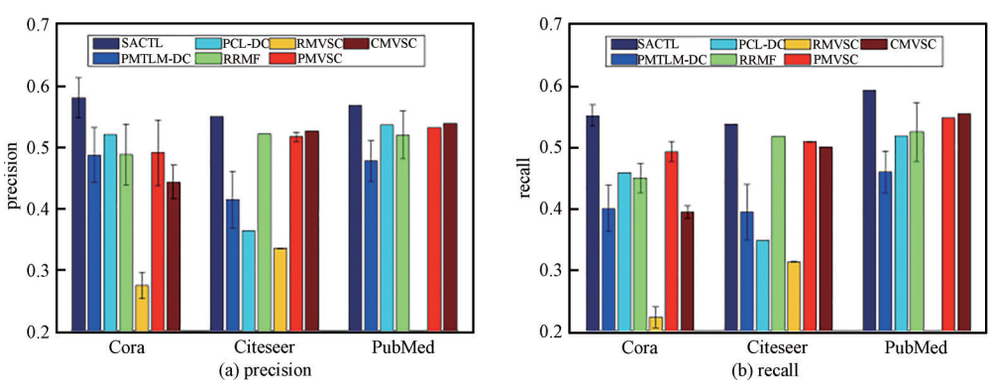

Fig.1 Performance comparison of different algorithms in precision and recall

图1 不同算法在精确度和召回率方面的性能比较

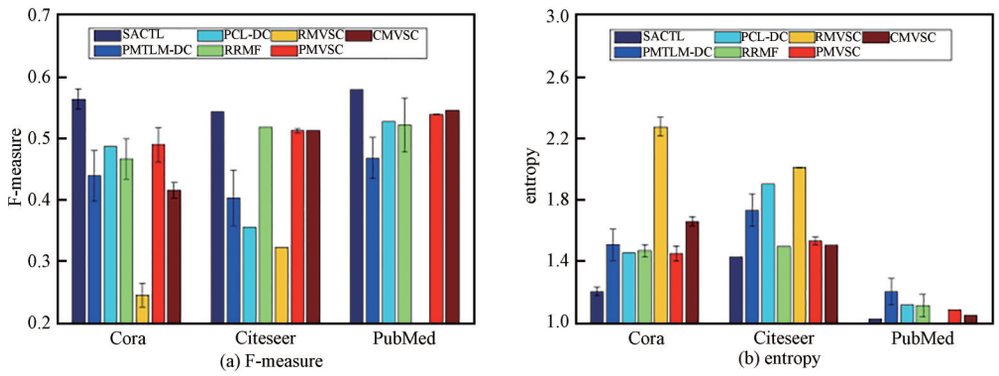

Fig.2 Performance comparison of different algorithms in F value and entropy
图2 不同算法在F值和熵方面的性能比较

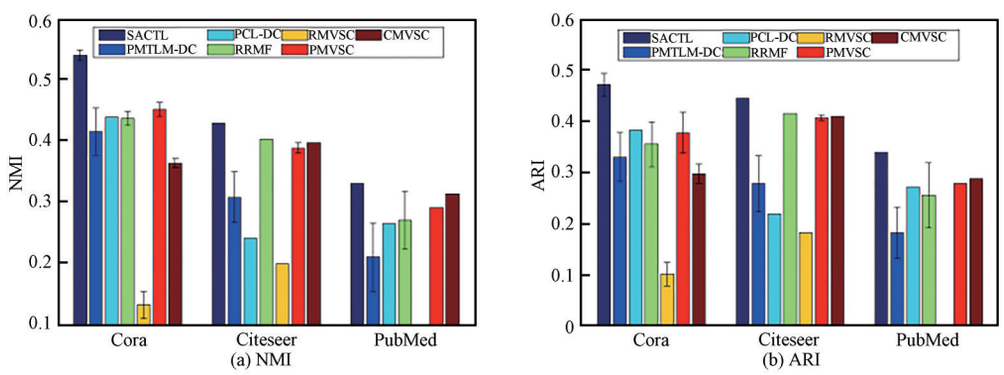

Fig.3 Performance comparison of different algorithms in NMI and ARI
图3 不同算法在NMI和ARI方面的性能比较

参考文献：

[1] 张雪松,贾彩燕. 一种基于频繁词集表示的新文本聚类方法[J]. 计算机研究与发展, 2018,55(1):102-112. (ZHANG Xuesong, JIA Caiyan. A new documents clustering method based on frequent itemsets[J]. Journal of Computer Research and Development, 2018,55(1):102-112.)

[2] 陈季梦,陈佳俊,刘杰,等. 基于结构相似度的大规模社交网络聚类算法[J]. 电子与信息学报, 2015,37(2):449-454. (CHEN Jimeng, CHEN Jiajun, LIU Jie, et al. Clustering algorithms for large-scale social networks based on structural similarity[J]. Journal of Electronics & Information Technology, 2015,37(2):449-454.)

[3] CHANG J, BLEI D M. Relational topic models for document networks[C]// In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics(AISTATS 2009). Clearwater Beach,Florida,USA:IEEE, 2009:81-88.

[4] COHN D A, HOFMANN T. The missing link-a probabilistic model of document content and hypertext connectivity[C]// Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, CO, USA: IEEE, 2015: 409-415.

[5] ZHU Y,YAN X,GETOOR L,et al. Scalable text and link analysis with mixed-topic link models[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago,IL,USA:IEEE, 2013:473-481.

[6] LI W,YEUNG D. Relation regularized matrix factorization[C]// Proceedings of the 21st International Joint Conference on Artificial Intelligence. California,USA:IEEE, 2019:1126-1131.

[7] 童绪军,吴义春. 数据挖掘中一种改进的谱组合聚类算法[J]. 太赫兹科学与电子信息学报, 2020,18(3):497-503. (TONG Xujun, WU Yichun. An improved spectral ensemble clustering algorithm in data mining[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(3):497-503.)

[8] KUMAR A, RAI P, DAUME H. Co-regularized multi-view spectral clustering[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada,Spain:IEEE, 2018:1413-1421.

[9] XIA R,PAN Y,DU L,et al. Robust multi-view spectral clustering via low-rank and sparse decomposition[C]// Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec,Canada:IEEE Press, 2014:2149-2155.

[10] YANG T,JIN R,CHI Y,et al. Combining link and content for community detection: a discriminative approach[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris,France:ACM, 2019:927-936.
[11] 邱云飞,刘畅. 基于加权集成Nystrom采样的谱聚类算法[J]. 模式识别与人工智能, 2019,32(5):420-428. (QIU Yunfei,LIU Chang. Spectral clustering algorithm based on weighted ensemble Nystrom sampling[J]. Pattern Recognition and Artificial Intelligence, 2019,32(5):420-428.)
[12] 李海洋,王恒远. 基于TL1范数约束的子空间聚类方法[J]. 电子与信息学报, 2017,39(10):2428-2436. (LI Haiyang,WANG Hengyuan. Subspace clustering method based on TL1 norm constraints[J]. Journal of Electronics & Information Technology, 2017,39(10):2428-2436.)
[13] 仰迪,白延琴,李倩. 半监督距离度量学习的内蕴加速投影梯度算法[J]. 运筹学学报, 2018,22(2):445-452. (YANG Di,BAI Yanqin, LI Qian. An intrinsic accelerated projection gradient algorithm for semi-supervised metric learning[J]. Operations Research Transactions, 2018,22(2):445-452.)
[14] DENOEUX T, LI S M, SRIBOONCHITTA S. Evaluating and comparing soft partitions: an approach based on Dempster-Shafer theory[J]. IEEE Transactions on Fuzzy Systems, 2018,26(3):1231-1244.

作者简介：

原 虹(1981-)，女，硕士，讲师，研究方向为计算机应用、数据挖掘.email:273564542@qq.com.

赵 丽(1973-)，女，硕士，副教授，研究方向为数据挖掘、机器学习技术.

王溢琴(1980-)，女，硕士，副教授，研究方向为数据挖掘、远程教育.

（上接第940页）

[8] SONG Hojin,NAGATSUMA Tadao. Present and future of terahertz communications[J]. IEEE Transactions on Terahertz Science and Technology, 2011,1(1):256-263.
[9] JARIWALA D, SANGWAN V K, LAUHON L J, et al. Carbon nanomaterials for electronics, optoelectronics, photovoltaics, and sensing[J]. Chemical Society Reviews, 2013,42(7):2824-2860.
[10] DRESSELHAUS M S,DRESSELHAUS G. Intercalation compounds of graphite[J]. Advances in Physics, 2002,33(51):228.
[11] KIM J T,CHUNG K H,CHOI C G. Thermo-optic mode extinction modulator based on graphene plasmonic waveguide[J]. Optics Express, 2013,21(13):15280-15286.
[12] YU Longhai, DAI Daoxin, HE Sailing. Graphene-based transparent flexible heat conductor for thermally tuning nanophotonic integrated devices[J]. Applied Physics Letters, 2014,105(25):251104-1-5.
[13] SENSALE-RODRIGUEZ B, YAN R, KELLY M M, et al. Broadband graphene terahertz modulators enabled by intraband transitions[J]. Nature Communications, 2012(3):780.
[14] LI Wei,CHEN Bigeng,MENG Chao,et al. Ultrafast all-optical graphene modulator[J]. Nano Letters, 2014,14(2):955-959.
[15] SAYEM A A,MAHDY M R C,JAHANGIR I,et al. Ultrathin ultra-broad band electro-absorption modulator based on few-layer graphene based anisotropic metamaterial[J]. Optics Communications, 2017(384):50-58.
[16] BERARDI S R,TIAN F,YAN R S,et al. Unique prospects for graphene-based terahertz modulators[J]. Applied Physics Letters, 2011,99(11):113104.
[17] HANSON G W. Dyadic green's functions and guided surface waves for a surface conductivity model of graphene[J]. Journal of Applied Physics, 2008,103(6):064302.

作者简介：

文 松(1987-)，男，硕士，高级工程师，主要研究方向为扩频通信.email:liang.alexandre@gmail.com.

赵其祥(1987-)，男，博士，讲师，主要研究方向为太赫兹科学、高功率真空器件、超材料.

马梦诗(1997-)，女，在读硕士研究生，主要研究方向为高功率太赫兹辐射源、模式转换器.

吕 游(1987-)，女，硕士，助理实验师，主要研究方向为太赫兹超材料、高功率太赫兹辐射源.

何国强(1990-)，男，博士，讲师，主要研究方向为毫米波/太赫兹传感技术、天线理论与技术、计算电磁学.