基于超图和样本自表征的谱聚类算法*

李永钢[1]，苏毅娟[2†]，何 威[1]，雷 聪[1]
(1. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004； 2. 广西师范学院 计算机与信息工程学院，南宁 530023)

摘 要： 针对传统谱聚类算法仅考虑数据点对点间的相互关系而未考虑数据间可能隐藏的复杂的相关性的问题，提出一种基于超图和自表征的谱聚类方法。首先，建立数据的超图，得到超图的拉普拉斯矩阵表示；然后利用 $l_{2,1}$-范数对样本进行行稀疏自表征，同时融入超图来描述数据间多层次的相互关系；最后，利用生成的自表征系数进行谱聚类。利用基于超图的样本自表征技术考虑了样本之间复杂的相关性。通过在 Hopkins155 等数据集上的实验表明，在聚类错误率评判标准下，算法优于现有基于普通图的谱聚类算法 SSC、SRC 等。

关键词： 谱聚类； 超图； 超图拉普拉斯； 样本自表征
中图分类号： TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2017)06-1621-05
doi:10.3969/j.issn.1001-3695.2017.06.005

Hypergraph and self-representation for spectral clustering

Li Yonggang[1], Su Yijuan[2†], He Wei[1], Lei Cong[1]
(1. *Guangxi Key Laboratory of Multi-source Information Mining & Security, Guangxi Normal University, Guilin Guangxi* 541004, *China*; 2. *College of Computer & Information Technology, Guangxi Teachers Education University, Nanning* 530023, *China*)

Abstract: To solve the issue that the traditional spectral clustering methods constructed the similarity matrix by only considering the pairwise relationship of the data but ignoring the complicated correlations among samples, this paper put forward a hypergraph and self-representation based spectral clustering method, called hypergraph and self-representation for spectral clustering (HGSR). Firstly, the algorithm constructed a hypergraph which fully considered the relations of samples to output the hypergraph Laplacian matrix. Secondly, it conducted row sparse self-representation for all samples by utilizing an $l_{2,1}$-norm regularizer, and also put hypergraph Laplacian into the regulation to guarantee the local structure of each sample. In this way, similar samples were clustered into same cluster. At last, it obtained an affinity matrix for conducting spectral clustering. By utilizing the hypergraph based self-representation, it considered the complicate relationships between the samples. The experimental results of Hopkins155 dataset and some image datasets show that the proposed method outperforms the LSR, SSC and LRR, in terms of the subspace clustering error.

Key words: spectral clustering; hypergraph; hypergraph Laplacian matrix; sample self-representation

0 引言

近年来，谱聚类[1~5]由于能够将样本聚类成任意形状的簇，使得簇内数据尽可能相似，不同簇之间的性质差异应尽可能大[6]，在机器学习、模式识别和计算机视觉等领域中得到了广泛的应用。谱聚类方法成功的关键在于利用样本的局部或全局信息，预先构建一个基于相似性(或距离)的关联矩阵(相似图)[7]。传统谱聚类方法仅依据点对点之间的相互关系建立相似图，本文称之为普通图，例如稀疏子空间聚类(sparse subspace clustering, SSC)[8]、低秩表征聚类(low rank representation, LRR)[9]、光滑表征聚类(smooth representation clustering, SRC)[10]。然而在很多现实问题中，人们需要关注的物体之间的关系远不止点对关系，而基于点对关系的普通图往往会忽略数据之间的多层次的关系，造成信息损失从而不利于聚类。数据间多层次的关系会包含对于聚类更加有用的信息(如仅考虑均值来聚类数据会忽略样本的分布信息，通过引入反映样本分布的标准差来聚类数据会更加准确)。因此，用普通图表征一组复杂的相关的物体是不完全的，需要构建包含多层信息的相似图，这是谱聚类面对的一大挑战。

现实中的数据往往有噪声和离群值，不利于数据分析。传统的谱聚类方法 SSC、LRR、LSR 和 SRC 等，均采用的是基于 F-norm 的自表征模型(特别地 LRR 和 LSR 算法采用 F-norm 作为正则项，缺乏稀疏性)，其对噪声和离群值敏感，不利于聚类，这是谱聚类面对的又一大挑战。

为了解决上述两大挑战，本文在模型中利用能够表示多层信息的超图[11]代替普通图，并且采用对离群值鲁棒的基于自表征的 $l_{2,1}$-norm 作为损失项构建谱聚类模型。算法首先建立超图，得到超图的拉普拉斯矩阵；然后对所有样本进行自表征，

收稿日期：2016-04-13； 修回日期：2016-06-01 基金项目：国家自然科学基金资助项目(61450001,61263035,61573270)；国家“973”计划资助项目(2013CB329404)；中国博士后科学基金资助项目(2015M570837)；广西自然科学基金资助项目(2012GXNSFGA060004,2015GXNSFCB139011,2015GXNSFAA139306)；广西研究生教育创新计划资助项目(YCSZ2016045,XYCSZ2017064)

作者简介：李永钢(1989-)，男，河北保定人，硕士，主要研究方向为数据挖掘、机器学习；苏毅娟(1976-)，女(通信作者)，广西桂林人，副教授，主要研究方向为机器学习和数据挖掘(574717541@qq.com)；何威(1989-)，男，河南商丘人，硕士，主要研究方向为数据挖掘、机器学习；雷聪(1991-)，男，湖北黄石人，硕士，主要研究方向为机器学习、数据挖掘.

并且利用 $l_{2,1}$-norm 对模型进行行稀疏约束; 接着,加入基于超图拉普拉斯的 trace-norm [12] 对模型进行正则化,得到全局最优的自表征系数矩阵。最后,利用谱聚类得到聚类结果。

本文算法相对传统的谱聚类算法的优势在于: a) 在自表征过程中,利用超图中的超边将更多层次的信息融入到模型中,解决普通图信息损失问题,特别地,由于超图拉普拉斯利用样本的高阶关系保持样本的局部结构,使得数据的高阶局部信息得到保持,有利于提高聚类效果; b) 算法采用 $l_{2,1}$-norm 作为损失项,通过控制自表征矩阵的行稀疏解决离群值的干扰,使模型具有更好的鲁棒性; c) 文中也为模型提出了一种迭代的方法高效地进行求解,这种基于超图和样本自表征的聚类算法简称为 HGSR (hypergraph and self-representation for spectral clustering) 算法。

1 相关理论

1.1 样本自表征

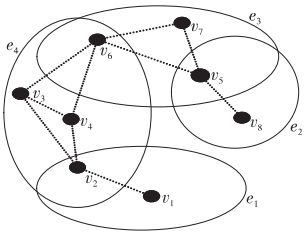
给定样本集 $X = [x_1, x_2, \cdots, x_n] \in \mathrm{R}^{d\times n}$ (d 为属性个数,n 为样本数) ,用 X 中的其他样本对 x_i ($i = 1, 2, \cdots, n$) $\in \mathrm{R}^{d\times 1}$ 进行线性表示的过程称为样本自表征 [10] 。根据样本自表征定义,需要找出一个列向量 $z_i \in \mathrm{R}^{n\times 1}$,使得 x_i 能够用 Xz_i 重新表示,其中 z_i 称为自表征系数。由于样本空间中往往会存在噪声或者离群点使得重新表示产生误差 e,即 $x_i = Xz_i + e$。因此,本文算法的目的在于找到最优自表征矩阵 $Z = [z_1, z_2, \cdots, z_n] \in \mathrm{R}^{n\times n}$,使得 X 与 XZ 之间的误差尽可能小。由于样本自表征系数依赖于全体样本而不仅仅依赖于单个样本,所以离群点的影响被降低,使得模型的鲁棒性得到增强。

1.2 超图

HGSR 算法将超图 [13] 定义成一个三元组 $G_H = (V, E, w)$,其中,$V = \{x_1, x_2, \cdots, x_n\}$ 表示样本点的集合; E 是 V 的非空子集,表示超边,代表了样本之间的层次关系; w 为超边的权重且是一个实数。作为普通图的概括,超图是一个自然的高层关系描述符 [14] 。与普通图的边仅连接(描述) 成对的样本点之间的关系不同,超图的边 e 可以包含任意大小的样本点的子集 [15,17] 。超边 e 中的样本点的个数 $\delta(e)$ 表示超边的度,而样本点 $v_i \in V$ 的度定义为 $d(v) = \sum_{v\in e, e\in E} w(e)$。为了更直观地介绍超图,本文给出一个简单超图的实例,如图 1 所示。图中的每一个顶点代表一部文献,每个超边表示与文献相应的作者,并用 $E = \{e_1 = \{v_1, v_2\}, e_2 = \{v_5, v_8\}, e_3 = \{v_5, v_6, v_7\}\}$ 表示超边的集合。为了更好地聚类数据,需要在原有关系基础上增加更多的关系,如文献发表的学术期刊名,这样一来每个学术期刊名就被认为是一条超边,用 $e_4 = \{v_2, v_3, v_4, v_6\}$ 表示。超图 G_H 可以用描述顶点与边之间关系的矩阵,即关联矩阵 $H \in \mathrm{R}^{|V|\times|E|}$ 表示,其中 $|V|$ 和 $|E|$ 分别表示样本数和超边数,H 中的元素定义如下:

$$h(v, e) = \begin{cases} 1 & \text{if}(v \in e) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

根据 H 的定义,有 $d(v) = \sum_{e\in E} w(e)\, h(v, e)$ 和 $\delta(e) = \sum_{v\in V} h(v, e)$ 成立。本文分别用 D_e 和 D_v 表示超边的度矩阵和点的度矩阵,并用 W_H 表示超边的权重矩阵。



(a)由 $V=\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ 组成的超图(虚线表示普通图)

	e_1	e_2	e_3	e_4
v_1	1	0	0	0
v_2	1	0	0	1
v_3	0	0	0	1
v_4	0	0	0	1
v_5	0	1	1	0
v_6	0	0	1	1
v_7	0	0	1	0
v_8	0	1	0	0

(b)超图(a)的关联矩阵 H

图 1 由 $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ 组成的超图及其关联矩阵

1.3 超图拉普拉斯矩阵

为了介绍超图拉普拉斯矩阵,本文先介绍普通图拉普拉斯矩阵,然后扩展到超图。

假设 G 表示由 n 个顶点组成的普通图。G 的邻接矩阵 $W \in \mathrm{R}^{n\times n}$ 定义为: 若第 i 个节点和第 j 个节点之间有边相连,则 $W_{ij} = 1$; 否则,$W_{ij} = 0$。顶点的度表示所有与该顶点相连的边的数目总和: $d_i = \sum_j W_{ij}$。定义对角矩阵 $D = \mathrm{diag}(d_1, \cdots, d_n)$,则普通图 G 的拉普拉斯矩阵为 $L = D - W$。本文将普通图拉普拉斯矩阵扩展到超图,称之为超图拉普拉斯矩阵。

尽管超图中的关联矩阵 H 完全可以描述超图的特性(H 中的元素表示点与超边之间的关系) 。但是,为了利用谱聚类算法进行聚类,需要得到样本与样本之间的表征关系,因此要为超图建立一个邻接矩阵和拉普拉斯矩阵 [16] 。一种可行的方法是: 利用相应超边的权重和基数的熵所加权的边来构造,也就是连通分量扩展和星形扩展 [17] 。另一种方法是对超图采用一个由邻接矩阵和相关联的拉普拉斯矩阵决定的矩阵表征,即规范化的拉普拉斯 [18] 。本文采用文献 [18] 中提到的方法来建立超图拉普拉斯。特别地,将一个超图的规范化的拉普拉斯矩阵的形式定义为 $\hat{L}_H = I_{|V|} - D_v^{-\frac{1}{2}} H W_H D_e^{-1} H^{\mathrm{T}} D_v^{-\frac{1}{2}}$,其中的 D_v 为对角点度矩阵,其对角元素 $d(v_i)$ 为矩阵 H 第 i 行之和; D_e 为对角边度矩阵,其对角元素为 $\delta(e_j)$,表示矩阵 H 的第 j 列之和。

2 算法描述

HGSR 算法将超图融入到样本自表征模型中,充分考虑了样本之间复杂的相互关系,保持了数据的高阶局部信息。通过 $l_{2,1}$-norm 和 trace-norm 对模型进行约束,确保具有相似表征系数的样本被聚类到一起。最后利用生成的自表征系数矩阵 Z 进行谱聚类。下文将详细说明算法是如何建立超图以及如何利用 HGSR 算法进行谱聚类分析。

2.1 超图的建立

对于超图,算法将数据集中的每个样本都认为是超图 $G_H = (V, E, w)$ 中的一个顶点,提出了通过链接样本以及其相关样本的方式生成超边 [19] 。特别地,每一个样本都可以被认为是一个响应向量,并且可以被其余的 $n-1$ 个样本的线性表示所估计,即 $x_i = Q_i z_i + \varepsilon_i$, $i = 1, 2, \cdots, n$。其中,$Q_i = [x_1, x_2, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_n]$ 表示除了 x_i 的所有样本(在其位置处用 0 代替) ; z_i 表示用其他样本近似表征 x_i 的表征系数, $\varepsilon_i \in \mathrm{R}^n$ 是表征误差项。一种原始的方法通过求解以下的问题得到 z_i 的稀疏解:

$$\min_{z_i} \| x_i - Q_i z_i \|_F + \lambda \| z_i \|_1 \tag{2}$$

其中: $\lambda > 0$ 是一个正则化参数,用来控制 z_i 的稀疏程度。当 λ

足够大时,l_1-norm 对损失项的惩罚将增大,使得不相关的样本对应的表征系数被强制设置成 0。如此,通过得到的表征系数能获得与 x_i 最相关的所有样本点的超边。

但是,最近的研究[20]表明,保存数据点局部几何结构信息比稀疏信息更有利于聚类。数据的局部几何结构更倾向于是一组数据的局部的近邻关系,这种近邻关系可以用每个样本的 K 近邻图(KNN 图)表示。直观地,相似的点应该具有相近的表征系数,结合超图,本文定义了一个基于超图的正则项,形式如下所示:

$$K(Z)=\frac{1}{2}\sum_{e\in E}\sum_{x_i,x_j\in V(e_i)}\frac{w(e)h(x_i,e)h(x_j,e)}{\delta(e)}\times\|z_i-z_j\|_2^2=$$
$$tr(Z\hat{L}_HZ^{\mathrm{T}})=tr(Z(I_{|V|}-D_v^{-\frac{1}{2}}HW_HD_e^{-1}H^{\mathrm{T}}D_v^{-\frac{1}{2}})Z^{\mathrm{T}}) \tag{3}$$

其中: W_H 为超边的权重矩阵; 本文将每个超边的权重 $w(e)$ 设置为 1; Z 为自表征系数矩阵; $H\in\mathrm{R}^{|V|\times|E|}$ 为超图的关联矩阵; $\hat{L}_H$ 为超图拉普拉斯矩阵; $V(e_i)$ 表示一组与超边 e_i 相关的数据点的集合; 正则项 $K(Z)$ 确保相似或者相近的样本 x_i、x_j 的表征系数 z_i、z_j 也相近。

2.2　HGSR 算法

由于 HGSR 算法是基于谱聚类模型的,在利用 HGSR 进行聚类分析之前先介绍谱聚类算法的步骤: a) 利用自表征系数构造样本集的关联矩阵; b) 通过计算关联矩阵前 k 个特征值与特征向量构建特征向量空间; c) 利用 K-means 算法对特征向量空间中的特征向量进行聚类。通过这种方法能够对数据集进行任意的聚类。

在进行聚类任务时,通常期望所得到的聚类模型应满足如下特性: 相似的样本其自表征系数也应当相似或者相近并且模型应该对离群的样本点和噪声鲁棒。传统的基于自表征的聚类方法模型采用 F-norm 作为损失项,并且利用普通图拉普拉斯矩阵作为正则项对自表征系数进行约束,得到最优自表征系数矩阵 Z^*,然后对 $(|Z^*|+|Z^{*\mathrm{T}}|)/2$ 采用谱聚类进行聚类分析。这类方法对离群样本点和噪声敏感,同时数据间相关信息有所损失不利于聚类。例如 LSR 中的聚类模型:

$$\min_Z J(Z)=\|X-XZ\|_F^2+\lambda tr(ZZ^{\mathrm{T}})=$$
$$\|X-XZ\|_F^2+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|z_i-z_j\|_2^2+\frac{1}{n}\|Z^{\mathrm{T}}e\|_2^2 \tag{4}$$

其中: e 为全 1 向量,可以看做是对所有的表征系数赋相同的权重,忽略了表征系数之间是否相近这一重要信息。鉴于 $l_{2,1}$-norm 对离群点的鲁棒性[21],本文用其作为损失项,并在模型中融入超图拉普拉斯矩阵对自表征系数 Z 进行约束(即式(3)),确保相似样本的表征系数也相近。最终得到如下的目标函数:

$$\min_Z J(Z)=\|X-XZ\|_{2,1}+\lambda tr(\hat{Z}L_HZ^{\mathrm{T}}) \tag{5}$$

其中: $\hat{L}_H$ 为超图 G_H 的拉普拉斯矩阵。因为式(5)的损失项不是二次的,所以式(5)可以使离群点比平方项 $\|X-XZ\|_F^2$ 有更小的重要性。通过利用基于超图的 trace-norm 作为正则项约束 Z,在模型中保持了数据的高阶局部信息,确保相似样本的表征系数也相近,最终提高聚类效果。算法 1 中描述了利用 HGSR 算法进行聚类的详细过程。

算法 1　HGSR 算法

输入: 训练样本 $X\in\mathrm{R}^{d\times n}$,正则化参数 λ。

输出: 聚类错误率。

a) 建立超图得到样本与超边的关系矩阵 H,然后利用 $\hat{L}_H=I_{|V|}-D_v^{\frac{1}{2}}HW_HD_e^{-1}H^{\mathrm{T}}D_v^{\frac{1}{2}}$ 得到超图的拉普拉斯矩阵 $\hat{L}_H$;

b) 利用算法 2 优化求解问题式(5)得到最优的表征系数矩阵 $Z^*\in\mathrm{R}^{n\times n}$;

c) 根据所得 Z^*,用谱聚类算法进行聚类;

d) 最后将聚类结果与原始类别对比,计算聚类错误率。

2.3　算法优化

由于目标函数式(5)是凸的且非光滑,难以直接求出最优解 Z^*。为此,本文提出一种高效的优化算法来求解目标函数。

首先,将目标函数式(5)关于 Z 的每一列 $z_i(1\leqslant i\leqslant n)$ 求导且令其等于 0,可以得到如下等式:

$$X^{\mathrm{T}}DXZ+Z(\alpha\hat{L}_H)+(-X^{\mathrm{T}}DX)=0 \tag{6}$$

因式(6)是标准 Sylvester 等式[21],所以算法采用 lyap 函数对其求解:

$$Z=\mathrm{lyap}(X^{\mathrm{T}}DX,(\lambda\hat{L}_H),(-X^{\mathrm{T}}DX)) \tag{7}$$

其中: X 和 $\hat{L}_H$ 已知,D 为对角阵,令 $U=X-XZ=[u_1,\cdots,u_n]^{\mathrm{T}}$,$D$ 的对角元素为 $d_{ii}=\frac{1}{\|u_i\|_2}$,注意到 D 是未知的且取决于 Z,因此提出使用一种迭代的方法对问题进行求解,即算法 2。

算法 2

输入: 数据集 $X\in\mathrm{R}^{d\times n}$,参数 λ。

输出: $Z\in\mathrm{R}^{n\times n}$。

a) **初始化:** $D^0\in\mathrm{R}^{d\times d}$,令 $k=0$;

b) **重复;**

c) **计算** $Z^{(k+1)}=\mathrm{lyap}(X^{\mathrm{T}}D^{(k)}X,(\lambda\hat{L}_H),(-X^{\mathrm{T}}D^{(k)}X))$;

计算对角矩阵 $D^{(k+1)}$,**其中令**

$U^{(k)}=X-XZ^{(k)}=[u_1^{(k)},\cdots,u_n^{(k)}]^{\mathrm{T}}$,$D^{(k)}$ **为对角阵,对角元素为** $d_{ii}^{(k)}=\frac{1}{\|u_i^{(k)}\|_2}$;

d) $k=k+1$;

e) **直到** $k+1$ **次目标函数与** k **次目标函数的差值小于** 10^{-5},**输出最优** Z^*。

定理 1　每一次迭代,算法 2 能够使得目标函数式(5)收敛。

证明　在算法 2 中,令 $S^{(k)}=tr(Z^{(k)}\hat{L}_H(Z^{(k)})^{\mathrm{T}})$ 表示第 k 次迭代时目标函数的正则项部分,$S^{(k)}$、$Z^{(k)}$ 分别表示第 k 次迭代时 S 和 Z 的优化值。根据文献[21]中的迭代再加权框架,优化非光滑项 $\|X-XZ\|_{2,1}$ 可以转换成迭代地更新 $\min_Z tr(U^{\mathrm{T}}DU)$ 中的 D 和 U,其中 $U=X-XZ$,D 为对角阵,对角元素为 $d_{ii}=\frac{1}{\|u_i\|_2}$,即

$$\lambda S^{(k+1)}+tr((U^{(k+1)})^{\mathrm{T}}D^{(k)}U^{(k+1)})\leqslant$$
$$\lambda S^{(k)}+tr((U^{(k)})^{\mathrm{T}}D^{(k)}U^{(k)}) \tag{8}$$

将 $tr(\cdot)$ 形式变换成和的形式,式(8)变成如下形式:

$$\lambda S^{(k+1)}+\sum_{i=1}^{n}\frac{\|u_i^{(k+1)}\|_2^2}{2\|u_i^{(k)}\|_2}\leqslant S^{(k)}+\sum_{i=1}^{n}\frac{\|u_i^{(k)}\|_2^2}{2\|u_i^{(k)}\|_2}$$

其中的 $u_i^{(k+1)}$ 表示 $U_i^{(k+1)}$ 的第 i 行,通过简单的变换可以得到

$$\lambda S^{(k+1)}+\sum_{i=1}^{n}\left(\frac{\|u_i^{(k+1)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k+1)}\|_2+\|u_i^{(k+1)}\|_2\right)\leqslant$$
$$\lambda S^{(k)}+\sum_{i=1}^{n}\left(\frac{\|u_i^{(k)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k)}\|_2+\|u_i^{(k)}\|_2\right) \tag{9}$$

通过简单地重组最终有

$$\sum_{i=1}^{n}\left(\frac{\|u_i^{(k+1)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k+1)}\|_2-\left(\frac{\|u_i^{(k)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k)}\|_2\right)\right)+\lambda S^{(k+1)}+\sum_{i=1}^{n}\|u_i^{(k+1)}\|_2\leqslant\lambda S^{(k)}+\sum_{i=1}^{n}\|u_i^{(k)}\|_2 \tag{10}$$

定理 2 对于任意非零行向量 $w\in\mathrm{R}^n$ 和 $w_0\in\mathrm{R}^n$ 有下式成立:

$$\|w\|_2-\frac{\|w\|_2^2}{2\|w_0\|_2}\leqslant\|w_0\|_2-\frac{\|w_0\|_2^2}{2\|w_0\|_2} \tag{11}$$

由定理 2 可知 $\frac{\|w\|_2^2}{2\|w_0\|_2}-\|w\|_2-\left(\frac{\|w_0\|_2^2}{2\|w_0\|_2}-\|w_0\|_2\right)\geqslant 0$ 成立。因此有

$$\lambda\sum_{i=1}^{n}\left(\frac{\|u_i^{(k+1)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k+1)}\|_2-\left(\frac{\|u_i^{(k)}\|_2^2}{2\|u_i^{(k)}\|_2}-\|u_i^{(k)}\|_2\right)\right)\geqslant 0 \tag{12}$$

根据式(10)和(12)可知

$$\lambda S^{(k+1)}+\sum_{i=1}^{n}\|u_i^{(k+1)}\|_2\leqslant\lambda S^{(k)}+\sum_{i=1}^{n}\|u_i^{(k)}\|_2 \tag{13}$$

成立,即 $\|X-XZ^{(k+1)}\|_{2,1}+\lambda tr(Z^{(k+1)}\hat{L}_H(Z^{(k+1)})^{\mathrm{T}})\leqslant\|X-XZ^{(k)}\|_{2,1}+\lambda tr(Z^{(k)}\hat{L}_H(Z^{(k)})^{\mathrm{T}})$ 成立。说明算法 2 能够使目标函数最终收敛。

3 实验分析

本文运用 HGSR 算法在运动分割和图像聚类两类应用中进行聚类,然后,将 HGSR 算法与目前聚类效果较好的基于普通图的聚类方法,如 LRR、LSR、SRC 和 SSC 算法进行对比。特别地,为了说明超图优于普通图,本文也采用了与式(5)相同的损失项和正则项,与基于普通图拉普拉斯矩阵的算法(称之为 GSR(graph and self-representation for spectral clustering))进行了对比。

3.1 实验设置和评价标准

本文算法通过 MATLAB 语言编程,且所有实验均在 Windows 7 系统下的 MATLAB 2014a 软件上运行测试。实验用到的数据集信息如下:

a) Hopkins155[22] 是一个运动分割数据集,包含 155 个经过抽取特征节点而生成的视频序列,每一个序列都是一个数据集,且每个序列含有(来自两个或三个) 39 ~ 550 个运动样本点。本文将原始数据用 PCA 降至 12 维,所有的算法在每一个序列上都进行了实验。对于每个算法,本文对所有序列采用相同的参数。

b) ORL[23] 是由剑桥 Olivetti 实验室提供的人脸数据集,包含来自 40 个人的共 400 张面部图像,本文用前 10 个人的共 100 张图像来进行实验。每张人脸图像被预处理成 16 × 16 像素大小的图片,即每张图片包含 256 个属性。

c) YaleB[10] 是标准的人脸数据集,噪声样本较多,这对聚类是一个大的挑战。数据集包含 38 个人在不同光照和姿势下的人脸图片。本文用前 10 个人的 64 张正面脸部图像共 640 个样本,将图像剪裁成 64 × 32 大小的图片进行实验。同样地用 PCA 将数据投影到 10 × 6 维。

d) Zoo[18] 是各种动物图像的数据集,包含 100 个动物、17 个属性(包括毛发、羽毛、卵、奶、腿、尾巴等等),本文将其聚类成七类。

与最近的大多数文献[9,10]一样,本文采用聚类错误率(clustering error,CE)作为聚类效果好坏的评价标准[16]。CE 能够在最优排序下通过匹配实验结果和样本真实值产生最小的误差,其形式定义如下:

$$CE=1-\frac{1}{N}\sum_{i=1}^{N}\varphi(Er_i,\mathrm{map}(Tl_i)) \tag{14}$$

其中: Er_i 和 Tl_i 分别表示第 i 个数据点的实验输出标签和样本真实标签; $\varphi(x,y)=1$ 当且仅当 $x=y$,其他情况 $\varphi(x,y)=0$; $\mathrm{map}(Tl_i)$ 是最优投影函数,通过 Kuhn-Munkres 算法[19] 可以有效地将聚类输出标签转换成与样本的真实标签相符合的形式。

3.2 实验结果与分析

为了说明所提出的优化算法(算法 2)的有效性以及高效性,在图 2 中,本文将算法 2 在各个数据集上的收敛性展示出来。由图 2 可以明显看出,随着迭代次数的增加,算法 2 能够快速收敛到全局最优解。由于设定目标函数小于 10^{-5} 时即表示收敛,所以在各个数据集上的迭代次数会有所不同。特别地,在 ORL 数据集上本文算法仅迭代 8 次就收敛到最优解,这也说明了算法 2 的高效性。

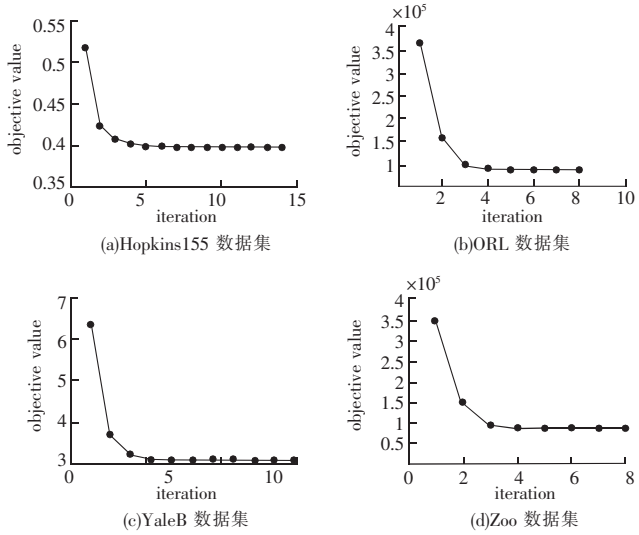


图 2 各数据集上,不同迭代次数时式(5)的目标函数值及收敛情况

在表 1 中,本文列出了在 Hopkins155 数据集上各算法所得到的聚类错误率。为了全面评价算法的有效性和高效性,本文从聚类错误率的最大值、均值、最小值以及标准差四个方面直观评价算法的性能。本文算法 HGSR 得到的聚类平均错误率为 2.48%,其他算法中所得到的最好结果为 SRC 的 3.35%。本文算法的标准差比其他算法都要小,说明了算法的稳定性。由于很多序列可以很容易地被聚类,所以,所有算法在这些序列上得到的最小聚类错误率为零。为了公平对比,本文亦将 HGSR 算法和具有相同重建误差项但采用普通图拉普拉斯矩阵的算法模型(简称 GSR)作了对比。由表 1 最后两列可以看出,本文算法在均值(mean)和标准差(STD)都优于采用普通图拉普拉斯矩阵的算法。

表 1 Hopkins155 上各算法所得聚类结果 /%

比较项	算法					
	SSC	LRR	LSR	SRC	GSR	HGSR
max	46.97	47.64	39.71	46.70	38.86	42.37
mean	3.92	5.14	4.21	4.24	3.35	2.48
min	0	0	0	0	0	0
STD	7.61	10.04	8.60	9.80	7.7	6.4

表 2 ~ 4 分别为各算法在 extended Yale Face B、ORL、Zoo

数据集上所得的聚类错误率,很明显可以看出本文的算法聚类效果均优于其他四种算法; 同样可以看出采用基于超图的算法 HGSR 比基于普通图的算法 GSR 聚类效果要好。由于采用 $l_{2,1}$-norm 作为损失项,GSR 比其余算法中聚类效果最好的算法 SRC(采用 F-norm) 聚类效果要好,这与 $l_{2,1}$-norm 比 F-norm 对离群值和噪声鲁棒是一致的。

表 2　Extended Yale Face B 上各算法所得聚类错误率对比

	算法					
	LRR	LSR	SSC	SRC	GSR	HGSR
CE/%	35.00	27.50	48.81	26.56	25.94	25.00

表 3　ORL 上各算法采用所得聚类错误率对比

	算法					
	LRR	LSR	SSC	SRC	GSR	HGSR
CE/%	53.75	22.25	22.50	21.25	19.50	17.25

表 4　Zoo 上所得聚类错误率对比

	算法					
	LRR	LSR	SSC	SRC	GSR	HGSR
CE/%	31.68	30.69	45.54	31.68	25.74	24.75

4　结束语

本文提出一种新的基于超图和样本自表征的聚类分析算法——HGSR 算法。HGSR 算法创新地使用了基于重构的行稀疏自表征和超图技术来对数据进行聚类分析。使用 $l_{2,1}$-norm 作为损失项,加强了模型的鲁棒性,同时用基于超图的迹范式作为正则项,保持了数据的高阶局部信息,确保了相似的样本表征系数也相近。通过以上技术充分考虑了样本之间更多的相互关系,很好地解决了传统聚类算法的信息损失和对离群值以及噪声敏感的问题。通过四个数据集对 HGSR 算法进行验证,并将 HGSR 算法和 SRC 等算法进行比较,鉴于聚类错误率的评价标准,实验结果表明,HGSR 算法比 SRC 等算法聚类效果要好。在未来的工作中,考虑将超图和自表征应用到高维数据的分析中,使得算法能够处理高维的大数据。

参考文献:

[1] 李俊英,汪西莉. 一种新的大规模复杂图像分割的谱聚类方法[J]. 计算机应用研究,2011,28(5):1994-1997.

[2] Zhu Xiaofeng,Huang Zi,Shen Hengtao,*et al*. Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition,2012,45(8):3003-3016.

[3] Zhu Xiaofeng,Li Xuelong,Zhang Shichao,*et al*. Robust joint graph sparse coding for unsupervised spectral feature selection [J]. IEEE Trans on Neural Networks & Learning Systems,2016,PP(99):1-13.

[4] Zhang Shichao,Jilian Zhang,Zhu Xiaofeng,*et al*. Missing value imputation based on data clustering [M]//Transactions on Computational Science I. Berlin: Springer-Verlag, 2008: 128-138.

[5] Lu Canyi,Min Hai,Zhao Zhongqiu,*et al*. Robust and efficient subspace segmentation via least squares regression [C]//Proc of the 12th European Conference on Computer Vision. Berlin: Springer, 2012: 347-360.

[6] Papa D A,Markov I L. Hypergraph partitioning and clustering [M]// Handbook of Approximation Algorithms and Mataheuristics. [S. l.]: Chaprman & Hall/CRC,2007.

[7] Zhu Xiaofeng,Huang Zi,Cui Jiangtao,*et al*. Video-to-shot tag propagation by graph sparse group Lasso [J]. IEEE Trans on Multimedia, 2013,15(3):633-646.

[8] Elhamifar E,Vidal R. Sparse subspace clustering: algorithm,theory, and applications [J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2012,35(11):2765-2781.

[9] Liu Guangcan,Lin Zhouchen,Yan Shuicheng,*et al*. Robust recovery of subspace structures by low-rank representation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2013,35(1):171-184.

[10] Hu Han,Lin Zhouchen,Feng Jiashi,*et al*. Smooth representation clustering [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press,2014: 3834-3841.

[11] Yu Jun,Tao Dacheng,Wang Meng. Adaptive hypergraph learning and its application in image classification [J]. IEEE Trans on Image Processing,2012,21(7):3262-3272.

[12] Zhu Xiaofeng,Huang Zi,Cheng Hong,*et al*. Sparse hashing for fast multimedia search [J]. ACM Trans on Information Systems,2013, 31(2):1-24.

[13] Agrrwal S,Branson K,Belongie S. Higher order learning with graphs [C]//Proc of the 23rd International Conference on Machine Learning. New York: ACM Press,2010: 17-24.

[14] Agrrwal S,Lim J,Zelnikmanor L,*et al*. Beyond pairwise clustering [C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 838-845.

[15] Purkait P,Chin T J,Ackermann H,*et al*. Clustering with hypergraphs: the case for large hyper-edges [C]//Proc of European Conference on Computer Vision. Berlin: Springer,2014: 672-687.

[16] Huang Sheng,Yang Dan,Ge Yongxin,*et al*. Discriminant hyper-Laplacian projections and its scalable extension for dimensionality reduction [J]. Neurocomputing,2015,173(part2):145-153.

[17] Zhu Xiaofeng,Suk H I,Lee S W,*et al*. Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification [J]. IEEE Trans on Biomed Engineering,2016,63(3): 607-618.

[18] Qin Yongsong,Zhang Shichao,Zhu Xiaofeng,*et al*. Semi-parametric optimization for missing data imputation [J]. Applied Intelligence, 2007,27(1):79-88.

[19] John W,Yang A Y,Arvind G,*et al*. Robust face recognition via sparse representation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence,2014,44(12):2368-2378.

[20] Zhu Xiaofeng,Huang Zi,Shen Hengtao,*et al*. Linear cross-modal hashing for efficient multimedia search [C]//Proc of the 21st ACM International Conference on Multimedia. New York: ACM Press,2013: 143-152.

[21] Zhu Xiaofeng,Zhang Lei,Huang Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Trans on Image Processing,2014,23(9):3737-3750.

[22] Zhang Shichao,Zhang Chengqi,Yang Qiang. Data preparation for data mining [J]. Applied Artificial Intelligence,2003,17(5-6):375-381.

[23] Zhang Shichao,Zhang Jilian,Zhu Xiaofeng. Missing data imputation by utilizing information within incomplete instances [J]. Journal of Systems and Software,2011,84(3):452-459.