基于自表征和群组效应的子空间聚类算法

苏毅娟[1]，李永钢[2+]，杨利锋[2]，孙 可[2]，罗 奘[2]

（1. 广西师范学院 计算机与信息工程学院，广西 南宁 530023；
2. 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）

摘 要：为解决目前聚类算法对噪声敏感和缺乏考虑样本间相关性等问题，提出一种充分考虑样本间相关性，使构造的关联矩阵保持子空间结构的子空间聚类算法。利用 $\ell_{2,1}$ －norm 对每个样本进行自表征；群组效应确保相近样本的自表征系数亦相近，生成块对角化的样本自表征系数矩阵；根据自表征矩阵得到关联矩阵，在谱聚类模型下实现数据聚类。在 Hopkins155 等数据集上的实验结果表明，在聚类错误率评判标准下，该算法优于现有经典子空间聚类算法 SRC、SSC 等。

关键词：子空间聚类；自表征；群组效应；谱聚类；关联矩阵
中图法分类号：TP181 **文献标识号**：A **文章编号**：1000-7024（2017）02-0534-05
doi：10.16208/j.issn1000-7024.2017.02.047

Self-representation and grouping effect for subspace clustering

SU Yi-juan[1], LI Yong-gang[2+], YANG Li-feng[2], SUN Ke[2], LUO Yan[2]

(1. College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530023, China;
2. Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China)

Abstract: To solve the issues that previous clustering methods are sensitive to noise and fail to consider the correlations among samples, a subspace clustering algorithm was proposed by taking the correlations among samples into account, so that the similarity matrix of the proposed clustering method preserved the structure of subspace. An $\ell_{2,1}$-norm was utilized to represent each sample by training samples. The grouping effect of the data was designed to ensure that the coefficient of close samples was similar, aiming at generating a diagonal block self-representation coefficient matrix. An affinity matrix was obtained for conducting spectral clustering. Experimental results on datasets such as Hopkins155 indicate that the proposed algorithm significantly outperforms the state-of-the-art methods, such as SRC and SSC.

Key words: subspace clustering; self-representation; grouping effect; spectral clustering; affinity matrix

0 引 言

近几年来，基于谱聚类[1-4]的子空间聚类[5]方法由于能够识别任意形状的样本子空间从而取得良好聚类效果，在机器学习等领域已得到广泛应用。谱聚类成功的关键在于利用样本的局部或全局信息构建了一个基于相似性的关联矩阵（affinity matrix）即相似图，因此如何建立相似图[6]对谱聚类尤为重要。目前比较流行基于表征的谱聚类算法，例如：稀疏子空间聚类（sparse subspace clustering: algorithm, theory and applications, SSC）[7]、低秩表征（low rank representation, LRR）[8]、光滑表征聚类（smooth representation clustering, SRC）[9]。前两种方法在数据信噪比小、子空间不相互独立时，其构造的块对角的关联矩阵的稀疏性或低秩性较差而不利于正确聚类。SRC 方法利用 F-norm 作为损失项构建的关联矩阵对噪声敏感。

因此，为了构造良好的关联矩阵进而获得更好的子空

收稿日期：2015-11-03；修订日期：2016-03-17
基金项目：国家自然科学基金项目（61450001、61263035、61363009、61573270）；国家 973 重点基础研究发展计划基金项目（2013CB329404）；中国博士后科学基金项目（2015M570837）；广西自然科学基金项目（2012GXNSFGA060004、2015GXNSFCB139011、2015GXNSFAA139306）；广西研究生教育创新计划基金项目（YCSZ2016045）
作者简介：苏毅娟（1976-），女，广西桂林人，副教授，研究方向为机器学习和数据挖掘；+通讯作者：李永钢（1989-），男，河北保定人，硕士，研究方向为数据挖掘、机器学习；杨利锋（1989-），男，广西桂林人，硕士，研究方向为数据挖掘、机器学习；孙可（1987-），男，河南永城人，硕士，研究方向为机器学习、数据挖掘；罗奘（1989-），男，安徽安庆人，硕士，研究方向为数据挖掘、机器学习。
E-mail: 574717541@qq.com

间聚类效果，本文首先从样本之间的相关性出发，对所有样本进行自表征，并通过 $\ell_{2,1}$-norm 和 trace-norm 分别对模型进行行稀疏和样本群组约束得到全局最优的自表征矩阵，并由之得到样本的关联矩阵。最后，利用谱聚类得到子空间聚类结果。在自表征过程中，用 trace-norm 的群组效应来确保每个样本都由与之自表征系数相近的样本表示，解决关联矩阵块对角结构性差的问题。而 $\ell_{2,1}$-norm 通过控制自表征矩阵的行稀疏解决噪音和离群点的干扰，使其具有更好的鲁棒性。为加强群组效应对关联矩阵的作用，文中亦提出了一种关联矩阵测量方法，实验结果表明，其聚类效果优于传统方法。本文将这种样本群组自表征聚类算法简称为 SRGE（self-representation and grouping effect for subspace clustering）。

1 相关理论

1.1 自表征

对于样本空间 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \boldsymbol{R}^{d\times n}$ 中的一个样本 $\boldsymbol{x}_i (i=1, 2\cdots, n) \in \boldsymbol{R}^{d\times 1}$，用 $\boldsymbol{X}$ 中的其它样本对 $\boldsymbol{x}_i$ 进行线性表示的过程称为样本自表征。由于样本自表征系数依赖于全体样本。因此，其对离群点鲁棒。

1.2 群组效应

群组效应在文献［10］中首次被提出：如果两个样本相似，那么它们的表征系数也应当彼此相近。群组效应定义如下：

定义 1 群组效应（grouping effect）：给定数据 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \boldsymbol{R}^{d\times n}$，对于 $\forall\, i\neq j$，如果 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\to 0$ 时有 $\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2\to 0$ 成立，其中 $\boldsymbol{z}_i$、$\boldsymbol{z}_j$ 分别为 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 的表征系数，则自表征矩阵 $\boldsymbol{Z}=[\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n] \in \boldsymbol{R}^{n\times n}$ 含有群组效应。

受 SRC 利用 trace-norm 的群组效应确保高度相关的样本被聚类到一起的启发，本文将群组效应融合到自表征模型中，以此生成块对角化的自表征系数矩阵 $\boldsymbol{Z}$，然后计算出关联矩阵 $\boldsymbol{J}$，最终提高聚类效果。

1.3 子空间聚类

给定数据集 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \boldsymbol{R}^{d\times n}$，其中 d 为属性个数，n 为样本数。假设这些样本点是分别从 k 个不同的子空间 $\{\boldsymbol{S}_i\}_{i=1}^{k}$（$i=1, \cdots, k$）里提取出来的，子空间聚类的目的就是将这些样本点正确地聚类到其所属的子空间。

目前基于谱聚类的子空间聚类算法的主要步骤是：首先，根据子空间策略构造样本集的关联矩阵 $\boldsymbol{J}$；然后，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。最后，利用 K-means 算法对特征向量空间中的特征向量进行聚类，从而实现子空间的聚类。如何构造块对角化的关联矩阵 $\boldsymbol{J}$ 划分子空间，使子空间内数据高度相似，不同子空间数据差异性大且成块对角，是谱聚类方法能否成功的关键。

2 算法描述

本文提出的 SRGE 算法通过充分利用样本之间的相关性来进行样本自表征，并通过 $\ell_{2,1}$-norm 和 trace-norm 分别对模型进行行稀疏和群组效应约束，由所得块对角化的自表征系数矩阵 $\boldsymbol{Z}$ 生成关联矩阵 $\boldsymbol{J}$，最后用谱聚类方法聚类。

根据样本自表征定义，需要找出一个列向量 $\boldsymbol{z}_i \in \boldsymbol{R}^{n\times 1}$，使得 $\boldsymbol{x}_i$ 能够用 $\boldsymbol{X}\boldsymbol{z}_i$ 重新表示。由于样本空间中往往会存在噪音或者离群点使得重新表示产生误差 $\boldsymbol{e}$，即 $\boldsymbol{x}_i=\boldsymbol{X}\boldsymbol{z}_i+\boldsymbol{e}$。因此，本文算法的目的在于找到最优自表征矩阵 $\boldsymbol{Z}=[\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n] \in \boldsymbol{R}^{n\times n}$，使得 $\boldsymbol{X}$ 与 $\boldsymbol{X}\boldsymbol{Z}$ 之间的误差尽可能小。现有的谱聚类方法[7-9]通过求解以下模型得到自表征矩阵 $\boldsymbol{Z}$

$$\min_{\boldsymbol{W}} \|\boldsymbol{X}-\boldsymbol{A}(\boldsymbol{X})\boldsymbol{Z}\|_l+\lambda Q(\boldsymbol{Z}) \quad \text{s.t. } \boldsymbol{Z}\in\boldsymbol{C} \tag{1}$$

其中，$\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \boldsymbol{R}^{d\times n}$，每一列为一个样本，$\boldsymbol{A}(\boldsymbol{X})$ 是一个字典矩阵，本文采用 $\boldsymbol{X}$ 代替，$\|\star\|_l$ 是损失函数，$Q(\boldsymbol{Z})$ 和 $\boldsymbol{C}$ 分别是对于 $\boldsymbol{Z}$ 的正则项和约束集，$\lambda>0$ 用来对损失项进行惩罚。

SRC 中采用 $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_F^2$ 作为损失项求解的 Z 不稀疏，且不能很好解决噪音和离群点干扰，因此我们采用 $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{2,1}$ 作为问题（1）中的损失项。本文的损失项不是二次的，使离群点会比平方项 $\|\boldsymbol{X}\boldsymbol{Z}-\boldsymbol{X}\|_F^2$ 有更小的重要性，并且利用具有群组效应的 trace-norm 作为正则项约束 $\boldsymbol{Z}$ 使其块对角化，最终提高聚类效果。即

$$Q(\boldsymbol{Z})=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2=\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T) \tag{2}$$

其中，$\boldsymbol{W}=(\boldsymbol{w}_{ij})$ 是权重矩阵用来衡量节点之间的亲密度，$\boldsymbol{L}$ 是拉普拉斯图矩阵：$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，其中 $\boldsymbol{D}$ 是对角矩阵：$\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{w}_{ij}$。常用的计算 $\boldsymbol{W}$ 的方法是采用热核函数或者 0－1 权重[10]来构造 k 近邻（k-nearest neighbor knn）图。实验中，本文采用 0－1 加权的方法来构造 knn 图（实验过程中根据经验 k 设置为 4）。当然，还有很多其它的子空间聚类方法来构造复杂的 knn 图，但这不是本文研究的重点。

为了防止数据之间的不稳定性，本文用 $\widetilde{\boldsymbol{L}}=\boldsymbol{L}+\xi\boldsymbol{I}$ 代替 $\boldsymbol{L}$（ξ 默认为 0.001），用 λ 作为调和参数防止模型过拟合，问题（1）转化为求解以下目标函数

$$\min_{\boldsymbol{Z}}\boldsymbol{J}(\boldsymbol{Z})=\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{2,1}+\lambda\mathrm{tr}(\widetilde{\boldsymbol{L}}\boldsymbol{Z}^T) \tag{3}$$

目标函数（3）不仅通过 trace-norm 确保每个样本都由与之具有相似表征系数（即强相关）的样本表示[11]，而且利用损失函数中的 $\ell_{2,1}$-norm 避免噪音和离群点的干扰，使其具有更好的鲁棒性。

传统方法利用 $\boldsymbol{J}_1=\frac{(|\boldsymbol{Z}^*|+|\boldsymbol{Z}^{*T}|)}{2}$ 求关联矩阵，然后利用谱聚类算法进行最终的聚类，这样得到的子空间

便具有同一子空间内的样本相似性高，不同子空间的样本差异性大，且所有子空间呈块对角化结构[12]的特征，很好解决了基于谱聚类的子空间聚类问题中构造一个良好关联矩阵的问题。

SRGE 算法的过程如下：

算法 1：SRGE 算法

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

输出：聚类错误率。

(1) 通过 knn 算法求出样本之间的 knn 图 $\boldsymbol{W}\in\boldsymbol{R}^{n\times n}$ 进而得到变形后的拉普拉斯矩阵 $\widetilde{\boldsymbol{L}}$；

(2) 利用 ADMM 算法，即算法 2 优化求解问题 (3) 得到最优的表征系数矩阵 $\boldsymbol{Z}^*\in\boldsymbol{R}^{n\times n}$；

(3) 利用式 (8) 和式 (9) 求得关联矩阵 $\boldsymbol{J}$；

(4) 根据所得 $\boldsymbol{J}$，用谱聚类算法将原始数据聚类成 m 个簇；

(5) 最后将聚类结果与原始类别做比较，计算聚类错误率。

3 算法优化

由于目标函数 (3) 是凸非光滑的，无法直接求解 $\boldsymbol{Z}$。为此，本文提出一种有效的优化算法来求解目标函数。

首先将目标函数式 (3) 对 $\boldsymbol{Z}$ 的每一列 $\boldsymbol{z}_i$ $(1\leqslant i\leqslant n)$ 求导且令其等于 0，可以得到如下等式

$$\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{Z}+\boldsymbol{Z}(\alpha\widetilde{\boldsymbol{L}})+(-\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})=0 \tag{4}$$

式 (4) 是一个标准的 Sylvester 等式[13]，可用 lyap 函数求得

$$\boldsymbol{Z}=\mathrm{lyap}(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X},(\lambda\widetilde{\boldsymbol{L}}),(-\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})) \tag{5}$$

其中，$\boldsymbol{X}$，$\widetilde{\boldsymbol{L}}$ 已知，$\boldsymbol{D}$ 为对角阵，令 $\boldsymbol{U}=\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n]^T$，$\boldsymbol{D}$ 的对角元素为 $\boldsymbol{d}_{ii}=\dfrac{1}{\|\boldsymbol{u}_i\|_2}$，注意 $\boldsymbol{D}$ 是未知的且取决于 $\boldsymbol{Z}$，本文用 ADMM 算法迭代的求解这一问题。

首先将式 (3) 分解成如下 N 个子问题：

$\arg\min\limits_{\boldsymbol{z}_i}\|\boldsymbol{x}_i-\boldsymbol{x}_i\boldsymbol{z}_i\|_{2,1}+\lambda\mathrm{tr}(\boldsymbol{z}_i\widetilde{\boldsymbol{L}}\boldsymbol{z}_i^T)$ 其中 $\boldsymbol{z}_i$ 为 $\boldsymbol{Z}$ 的列子向量且 $\mathrm{vec}(\boldsymbol{Z})=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots\cdots,\boldsymbol{z}_n]^T$。

由于式 (3) 采用的参数矢量迹范数是受约束的，使得 $\boldsymbol{Z}$ 中的元素不能独立处理，因此公式中软阈值 λ 的应用是不平凡的并且是低效率的。但是可将 λ 用在优化中，称作乘数交替方向法。采用加入虚变量的方式，可以使目标转化为如下形式

$$\begin{aligned}\arg\min_{\boldsymbol{Z},\boldsymbol{V}}&\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{2,1}+\lambda\mathrm{tr}(\boldsymbol{V}\widetilde{\boldsymbol{L}}\boldsymbol{V}^T)+\\&\rho\|\boldsymbol{Z}-\boldsymbol{V}\|_F^2\quad\text{s.t.}\quad\boldsymbol{Z}=\boldsymbol{V}\end{aligned} \tag{6}$$

由于目标函数式 (6) 形式复杂，所以本文利用扩展拉格朗日函数将式 (6) 转化成如下模型

$$\begin{aligned}L(\boldsymbol{Z},\boldsymbol{C},\boldsymbol{\Lambda})=&\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{2,1}+\lambda\mathrm{tr}(\boldsymbol{V}\widetilde{\boldsymbol{L}}\boldsymbol{V}^T)+\\&\frac{\rho}{2}\|\boldsymbol{Z}-\boldsymbol{V}\|_F^2+\mathrm{vec}(\boldsymbol{\Lambda})\mathrm{vec}(\boldsymbol{Z}-\boldsymbol{V})\end{aligned} \tag{7}$$

ADMM 算法的基本思想包括如下迭代步骤：

(1) $\boldsymbol{Z}^{(k+1)}=\arg\min\limits_{\boldsymbol{Z}}L(\boldsymbol{Z},\boldsymbol{V}^{(k)},\boldsymbol{\Lambda}^{(k)})$；

(2) $\boldsymbol{V}^{(k+1)}=\arg\min\limits_{\boldsymbol{V}}L(\boldsymbol{Z}^{(k+1)},\boldsymbol{V},\boldsymbol{\Lambda})$；

(3) $\boldsymbol{\Lambda}^{(k+1)}\leftarrow\boldsymbol{\Lambda}^{(k)}+\rho(\boldsymbol{Z}+\boldsymbol{V})$。

对于输入 $\boldsymbol{\Lambda}$ 和 ρ，算法的关键就是解决式 (6) 的最优解问题。根据 ADMM 算法，式 (3) 可被拆分成如下两个子问题：

第一个子问题：假如仅优化式 (6) 中的 $\boldsymbol{Z}$，当迹范数惩罚 $\mathrm{tr}(\boldsymbol{V}\widetilde{\boldsymbol{L}}\boldsymbol{V}^T)$ 使 $\boldsymbol{Z}$ 从目标函数中消失时，该子问题将被转换成简单最小二乘回归问题。第二个子问题：假如仅优化式 (6) 中的 $\boldsymbol{V}$，则当损失项 $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{2,1}$ 消失时，将允许 $\boldsymbol{V}$ 独自求解。通过将目标函数转化成这两个子问题，软阈值 λ 可得到有效利用。然后，将当前所估计的 $\boldsymbol{Z}$ 和 $\boldsymbol{V}$ 与 ADMM 算法的第三步结合，可更新拉格朗日乘子矩阵 $\boldsymbol{\Lambda}$ 的当前估计。目标函数中的惩罚参数 ρ 具有特殊作用：可利用有缺的估计 $\boldsymbol{\Lambda}$ 求解 $\boldsymbol{Z}$ 和 $\boldsymbol{V}$。

用 ADMM 算法求解最优解 $\boldsymbol{Z}^*$ 的伪代码如下：

算法 2：ADMM 算法

输入：数据集 $\boldsymbol{X}$，和惩罚参数 ρ；

输入/出：$\boldsymbol{Z}\in\boldsymbol{R}^{n\times n}$，$\boldsymbol{\Lambda}$。

(1) **初始化**：$\boldsymbol{Z}^0$，$\boldsymbol{V}^0$，$\boldsymbol{\Lambda}^0$

(2) **重复**

(3) $\boldsymbol{V}^{(k+1)}\leftarrow\boldsymbol{V}^{(k)}$；将 $\boldsymbol{Z}^{(k)}$ 带入 $\boldsymbol{Z}^{k+1}=lyap(\boldsymbol{X}^T\boldsymbol{D}^{(k)}\boldsymbol{X},(\lambda\widetilde{\boldsymbol{L}}),(-\boldsymbol{X}^T\boldsymbol{D}^{(k)}\boldsymbol{X}))$ 更新 $\boldsymbol{Z}$，其中令 $\boldsymbol{U}^{(k)}=\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(k)}=[\boldsymbol{u}_1^{(k)},\cdots,\boldsymbol{u}_n^{(k)}]^{\mathrm{T}}$，$\boldsymbol{D}^{(k)}$ 为对角阵，对角元素为 $\boldsymbol{d}_{ii}^{(k)}=\dfrac{1}{\|\boldsymbol{u}_i^{(k)}\|_2}$

(4) $\boldsymbol{\Lambda}^{(k+1)}\leftarrow\boldsymbol{\Lambda}^{(k)}+\rho(\boldsymbol{Z}+\boldsymbol{V})$

(5) $k=k+1$

(6) **直到** $\boldsymbol{Z}$ **最优，输出最优解** $\boldsymbol{Z}^*$

4 利用 SRGE 进行子空间聚类

利用上述方法得到最优的自表征矩阵 $\boldsymbol{Z}^*$，计算下式得到关联矩阵 $\boldsymbol{J}_1$

$$\boldsymbol{J}_1=\frac{(|\boldsymbol{Z}^*|+|\boldsymbol{Z}^{*T}|)}{2} \tag{8}$$

然后正如 SSC，LRR 和 LSR 中所用到的一样，利用谱聚类算法[14]来产生最后的聚类结果。

$\boldsymbol{J}_1$ 的有效性主要来源于 $\boldsymbol{Z}^*$ 的块对角性质，并没有用到群组效应的良好性质，因此，为了利用群组效应，我们采用新的关联矩阵，如下所示

$$\boldsymbol{J}_2=\left(\left|\frac{\boldsymbol{z}_i^{*T}\boldsymbol{z}_j^*}{\|\boldsymbol{x}_i\|_2\|\boldsymbol{x}_j\|_2}\right|^{\gamma}\right) \tag{9}$$

用 $\gamma>0$ 来控制样本间相似度的差异变动，γ 默认值为 2[9]。新的关联矩阵计算方法，可以看成用样本原始特征的

范数所标准化后新的表征向量的内积。这样的标准化防止了相似度的测量由于原始特征的振动（在运动分割问题和手写数字识别问题中尤为明显）而产生偏差。

5 实验分析

本文在 4 类应用中运用 SRGE 算法进行子空间聚类：运动分割，人脸图像聚类，手写数字图像聚类和心理学平衡聚类。随后将我们的算法与目前聚类效果较好的基于表征重建的方法对比，例如 SSC，LRR，LSR，SRC 算法。

5.1 实验数据集和评判标准

本文实验是在 win 7 系统下的 matlab2014a 软件上进行编程实验。实验用到的数据集：

Hopkins155[15]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，和很多文献中相同，对于每个算法，我们对所有序列采用相同的参数。Jaffe[16]国际通用的标准人脸数据集，其共有 10 位女性正脸面部表情的 213 张图像，每一张图片包含 32×32 个像素。USPS[17]是包含 9298 张图片的手写数字数据集，每一张图片包含 16×16 个像素。我们用每个数字的前 100 张来进行实验。Balance[18]是模拟心理学实验结果所产生的，共包含 625 个样本，每个样本包含 4 维属性。

本文采用聚类错误率（clustering error，CE）来评判准确度[9]。CE 是在最优排列下通过匹配结果和样本真值产生最小的误差，其形式定义如下

$$CE = 1 - \frac{1}{N}\sum_{i=1}^{N}\delta(p_i, map(q_i)) \tag{10}$$

其中，q_i 和 p_i 表示第 i 个数据点的输出标签和样本真实标签；$\delta(x, y) = 1$ 当且仅当 $x = y$，其它情况 $\delta(x, y) = 0$；$map(q_i)$ 是最优投影函数，将聚类标签转换成和样本真实标签相符合的形式，通过 Kuhn-Munkres 算法[19]可有效的进行计算。

5.2 实验结果与分析

为了公平的对比，我们对所有的算法采用相同的重建误差项。表 1 列出了 5 个方法在 Hopkins155 数据集上采用经典关联矩阵测量方法（8）得到的运动分割错误率。我们算法得到的聚类平均错误率为 3.35%，而其它算法中所得到的最好结果为 SSC 的 3.92%。可以看到表 1 中的数据结果与 CASS[1]中的有所不同，因为它们计算 CE 用的是近似求解。由于很多序列可以很容易的被分割，因此，所有算法在这些序列上得到的最小聚类错误率为零。算法所用的时间也在表 1 中一同列出，SRGE 比 LSR 慢，但是比 SSC 和 LRR 速度要快很多，由于对自表征矩阵 Z 的求解方法与 SRC 相似，所以 SRGE 和 SRC 算法所用时间基本相同。

表 2 和表 3 分别说明了算法 SRGE 采用传统的关联矩阵度量方法式（8）中的 J_1 在 USPS 和 Balance 数据集上的聚类错误率。实验结果表明，我们的算法的聚类效果明显的优于其它算法。表 4 中，Jaffe 数据集中包含 1 组中性表情数据（即噪声样本），其它聚类方法由于对噪声敏感聚类效果较差，而 SRGE 算法对噪声鲁棒，聚类错误率低于 1%。

表 1 Hopkins 155 上各算法采用 J_1 所得聚类结果/%

method	SSC	LRR	LSR	SRC	SRGE
Max	46.97	47.64	39.71	46.70	**38.86**
Mean	3.92	5.14	4.21	4.24	**3.35**
Min	0	0	0	0	**0**
Median	0	0.53	0.52	0.29	**0**
STD	**7.61**	10.04	8.60	9.80	7.7
Time/s	2.50	2.03	**0.12**	0.40	0.39

表 2 USPS 上各算法采用 J_1 所得聚类错误率对比

method	LRR	LSR	SSC	SRC	SRGE
CE/%	22.60	26.10	43.10	12.70	**12.10**

表 3 USPS 上各算法采用 J_2 所得聚类错误率对比

method	LRR	LSR	SSC	SRC	SRGE
CE/%	17.50	18.40	42.20	11.30	**11.00**

表 4 Jaffe 上各算法采用 J_1 所得聚类错误率对比

method	LRR	LSR	SSC	SRC	SRGE
CE/%	47.33	37.91	13.15	4.69	**0.94**

为了说明本文提出的关联矩阵的有效性，在 USPS 数据集上我们对每个算法都采用式（9）中的 J_2 进行聚类，结果在表 5 中列出，由表 2 和表 5 对比可看出：所有数据集上，各类算法利用 J_2 取得的聚类效果均优于利用 J_1 取得的聚类效果。

表 5 Balance 上各算法 J_1 所得聚类错误率对比

method	LRR	LSR	SSC	SRC	SRGE
CE/%	46.40	38.24	43.68	41.76	**35.20**

在图 1 和图 2 中我们分别将 USPS 中每个数字的前 5

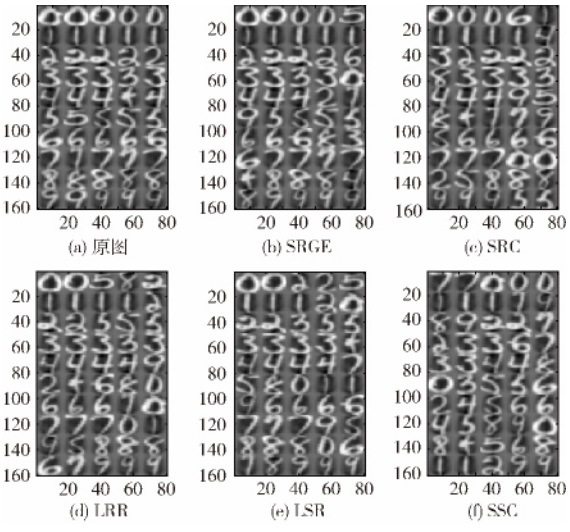

图 1 USPS 前 50 张图片各算法的聚类效果

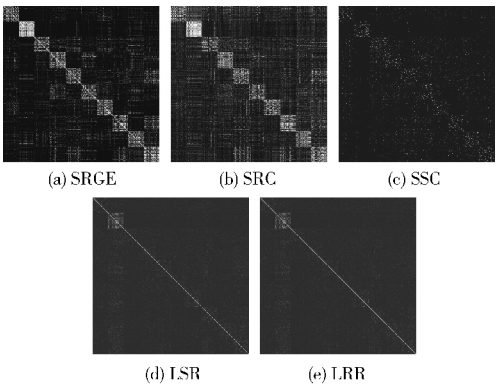

图 2 自表征矩阵 $Z*$ 的块对角化效果对比

张照片用 SRGE 等算法得到的聚类效果和所得自表征矩阵 $Z*$ 的对角化效果进行了展示。

图 1 中由于聚类分组使得各个数字的标签随机生成，我们将所得结果按照数字顺序排序并最终生成图像。图 1 中的第一行至第十行分别代表数字 0～10。SRGE 算法的聚类错误率为 26%，比其他算法中聚类效果最好的 SRC 的 36%要低。

图 2 中由于 USPS 数据集中的数据由 0～10 组成（即表示 10 个相互独立的子空间），因此，图 2 中每张图中的主对角线上均有 10 个方块，代表 10 个子空间。SRGE 算法得到的自表征矩阵 $Z*$ 的块对角化效果比其他算法得到的自表征矩阵块对角化效果更加明显，这也正是 J_1 有效性的由来。

6 结束语

本文提出一种聚类分析算法——SRGE 算法，即通过利用数据本身的相似性来构建自表征系数矩阵，并整合迹范数作为正则项，来达到有效聚类数据的目的。通过以上技术充分考虑了样本之间的相似性和群组效应，很好的解决了噪声污染问题。目标函数（3）虽然是凸的但却是非光滑的，不易直接进行求解，因此本文采用 ADMM 算法迭代的对其求解。

通过采用 4 个数据集对 SRGE 算法进行验证，并将 SRGE 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，SRGE 算法比 SRC 等算法效果要好。

参考文献：

[1] LU C，LIN Z，YAN S. Correlation adaptive subspace segmentation by trace lasso [C] //IEEE International Conference on Computer Vision. Sydney：IEEE，2013：1345-1352.

[2] WANG Huiqing，CHEN Junjie. Research of spectral clustering based on graph partition [J]. Computer Engineering and Design，2011，32 (1)：289-292 (in Chinese). [王会青，陈俊杰. 基于图划分的谱聚类方法研究 [J]. 计算机工程与设计，2011，32 (1)：289-292].

[3] Patel VM，Hien Van Nguyen，Vidal R. Latent space sparse subspace clustering [C] //IEEE International Conference on Computer Vision，2013：225-232.

[4] LU CY，MIN H，ZHAO ZQ，et al. Robust and efficient subspace segmentation via east squares regression [C] //Proceedings of the 12th European Conference on Computer Vision-Volume Part VII，2012：347-360.

[5] Vidal R. Subspace clustering [J]. IEEE Signal Processing Magazine，2011，28 (2)：52-68.

[6] Peng X，Zhang L，Yi Z. Scalable sparse sub-space clustering [C] //IEEE Conference on Computer Vision and Pattern Recognition，2013：430-437.

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

[8] Liu G，Lin Z，Yan S，et al. Robust recovery of subspace structures by low-rank representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35 (1)：171-184.

[9] Hu H，Lin Z，Feng J，et al. Smooth representation clustering [C] //IEEE Conference on Computer Vision and Pattern Recognition，2014：3834-3841.

[10] Xu Y，Zhong A，Yang J，et al. LPP solution schemes for use with face recognition [J]. Pattern Recognition，2010，43 (12)：4165-4176.

[11] Grave E，Obozinski GR，Bach FR. Trace lasso：A trace norm regularization for correlated designs [C] //Advances in Neural Information Processing Systems，2011：2187-2195.

[12] Feng J，Lin Z，Xu H，et al. Robust subspace segmentation with block-diagonal prior [C] //IEEE Conference on Computer Vision and Pattern Recognition，2014：3818-3825.

[13] Zhu X，Zhang L，Huang Z. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Transactions on Image Processing，2014，23 (9)：3737-3750.

[14] Chen WY，Song Y，Bai H，et al. Parallel spectral clustering in distributed systems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2011，33 (3)：568-586.

[15] Delong A，Osokin A，Isack HN，et al. Fast approximate energy minimization with label costs [J]. International Journal of Computer Vision，2012，96 (1)：1-27.

[16] Huang D，Shan C，Ardabilian M，et al. Local binary patterns and its application to facial image analysis：A survey [J]. IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)，2011，41 (6)：765-781.

[17] Huang GB，Zhou H，Ding X，et al. Extreme learning machine for regression and multiclass classification [J]. IEEE Transactions on Systems Man and Cybernetics，Part B：Cybernetics，2012，42 (2)：513-529.

[18] Pavao SL，Barbosa KAF，Tatiana de Oliveira Sato，et al. Functional balance and gross motor function in children with cerebral palsy [J]. Research in Developmental Disabilities，2014，35 (10)：2278-2283.

[19] Zhu X，Suk HI，Shen D. A novel matrixsimilarity based loss function for joint regression and classification in AD diagnosis [J]. Neuro Image，2014，100：91-105.