软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(12):3347−3357 [doi: 10.13328/j.cnki.jos.005235]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于ℓ_2范数的加权低秩子空间聚类*

傅文进, 吴小俊

(江南大学 物联网工程学院,江苏 无锡 214122)

通讯作者: 吴小俊, E-mail: wu_xiaojun@jiangnan.edu.cn

摘 要: 针对稀疏子空间聚类和最小二乘回归子空间聚类求得的表示系数存在类内过于稀疏和类间过于稠密的问题,利用ℓ_2范数,提出一种基于欧氏距离的且具有组效应的加权低秩子空间聚类算法,该算法通过基于欧氏距离的加权方式,使得最终的表示系数在保证同一子空间数据点联系的同时,减小不同子空间数据点之间的联系.利用该表示系数建立相似矩阵 J,将 J 应用到谱聚类得到聚类结果.实验结果表明,与当前流行的算法比较,该算法取得了较好的聚类效果.

关键词: 低秩;子空间聚类;组效应;ℓ_2范数;加权方式;谱聚类

中图法分类号: TP309

中文引用格式: 傅文进,吴小俊.基于ℓ_2范数的加权低秩子空间聚类.软件学报,2017,28(12):3347−3357. http://www.jos.org.cn/1000-9825/5235.htm

英文引用格式: Fu WJ, Wu XJ. Weighted low rank subspace clustering based on ℓ_2 norm. Ruan Jian Xue Bao/Journal of Software, 2017,28(12):3347−3357 (in Chinese). http://www.jos.org.cn/1000-9825/5235.htm

Weighted Low Rank Subspace Clustering Based on ℓ_2 Norm

FU Wen-Jin, WU Xiao-Jun

(School of Internet of things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: In order to solve the problem of over-sparsity for within-class coefficients and over-density for between-class coefficients in SSC and LSR, this paper proposes a new subspace clustering based on Euclidean distance using ℓ_2 norm. Using the weighted method based on Euclidean distance, the coefficient representation obtained by this algorithm maintains the connections of the data points from the same subspace. Meanwhile, the algorithm can eliminate the connections between clusters. The clusters can be produced by using the spectral clustering with the similarity matrix which is constructed by this coefficient representation. The results of experiments indicate the presented method improves the accuracy of clustering.

Key words: low rank; subspace clustering; grouping effect; ℓ_2 norm; weighted method; spectral clustering

随着科学技术的发展,聚类分析在计算机视觉和模式识别中扮演着越来越重要的角色.当前,高维数据成为互联网时代的主流,传统的聚类算法无法取得令人满意的聚类效果.如何有效处理高维数据的聚类问题,仍是当前聚类分析中的研究热点.近期,子空间聚类算法作为一种有效处理高维数据的工具被广泛应用在图像分割、模式识别、人工智能等领域.所谓子空间聚类,即要求将数据集中的数据点分配到其本质的所属低维子空间中.如对于具有多个人脸对象的图像集,不同的人脸对应着不同的子空间,为了识别出不同的人脸图像,需要对多个

* 基金项目: 国家自然科学基金(61373055, 61672265); 江苏省教育厅科技成果产业化推进项目(JH10-28)

Foundation item: National Natural Science Foundation of China (61373055, 61672265), Industry Project of Provincial Department of Education of Jiangsu Province (JH10-28)

收稿时间: 2016-02-28; 修改时间: 2016-08-10; 采用时间: 2016-12-19; jos 在线出版时间: 2017-03-24

CNKI 网络优先出版: 2017-03-24 15:31:40, http://kns.cnki.net/kcms/detail/11.2560.TP.20170324.1531.004.html

子空间的数据进行聚类.

定义 1(子空间聚类(subspace clustering,简称 SC))[1]. 给定一组数据 $X=[x_1\ x_2\ \ldots x_n]\in\mathbb{R}^{d\times n}$,设这组数据属于 K (K 已知或未知)个线性子空间 $\{S_i\}_{i=1}^{K}$ 的并,子空间聚类是指将这组数据分割为不同的类,在理想情况下,每一类对应不同的子空间.

当前,子空间聚类算法主要有代数法[2]、统计法[3]、迭代法[4]和基于谱聚类[5,6]的算法.基于谱聚类的子空间聚类算法能够有效处理高维数据的聚类问题,其算法核心在于构建一个具有子空间结构性质的相似矩阵.目前,有两种建立相似矩阵的方式:基于距离的方法[7,8]和基于线性表示的方法[9−13].基于线性表示的方法由于其高效的聚类性能及其对数据集中的噪声和离散点的鲁棒性,成为该领域的研究热点.

本文在充分研究基于线性表示的子空间聚类算法的基础上,受到 LLC(locality-constrained linear coding)[14]的启发,将加权重构正则项引入对称低秩表示子间聚类算法(symmetric low-rank representation,简称 SLRR)[13]的优化问题中,提出一种基于 ℓ_2 范数的加权低秩子空间聚类方法.此方法在保留数据点间全局结构的同时,突出数据点之间的局部特性,即:在保证同一子空间内数据点之间联系的同时,减小不同子空间数据点之间的联系.我们将此方法命名为从全局结构到局部结构的子空间聚类方法(global structure and local structure of data for subspace clustering,简称 GLSC).

1 相关原理

1.1 基于线性表示的子空间聚类相关原理及方法

基于线性表示的子空间聚类算法利用整个数据矩阵作为字典,求取每个数据点的线性表示,然后利用得到的系数矩阵建立相似矩阵,在高维数据聚类中,这些方法取得较好的效果.它们通过解公式(1)得到原数据集的表示系数 Z^*:

$$\begin{aligned}&\min_{Z}\ a\,\|X-A(X)Z\|_l+\Omega(X,Z)\\&\text{s.t.}\ \ Z\in\mathcal{C}\end{aligned}\tag{1}$$

其中,$X\in\mathbb{R}^{d\times n}$ 是数据矩阵,每一列表示一个数据样本;$A(X)$表示字典,一般地,$A(X)=X$;$\|\cdot\|_l$ 表示某种范数,如 ℓ_1,ℓ_2 或者 $\ell_{2,1}$;$\Omega(X,Z)$和$\mathcal{C}$分别表示正则项和约束集合;α表示惩罚因子,用于平衡两项[9].

- 当$\|\cdot\|_l$ 为 $\|\cdot\|_F^2$,$\Omega(X,Z)=\|Z\|_1$ 时,为稀疏子空间聚类算法(sparse subspace clustering,简称 SSC)[10];
- 当$\|\cdot\|_l$ 为$\|\cdot\|_{2,1}$,$\Omega(X,Z)=\|Z\|_*$时,为低秩子空间聚类算法(low-rank representation for subspace clustering,简称 LRR)[11];
- 当$\|\cdot\|_l$ 为 $\|\cdot\|_F^2$,$\Omega(X,Z)=\|Z\|_F^2$ 时,为最小二乘回归子空间聚类算法(subspace segmentation via least squares regression,简称 LSR)[12].

其中,SSC 求得的表示系数虽然消除了不同子空间数据点之间的联系,但是同一子空间内的数据点之间的联系过于稀疏.并且,SSC 不具有组效应特性[12].LRR 和 LSR 具有组效应特性,但是 LRR 时间复杂度较高,每次迭代都要对数据矩阵进行奇异值分解.LSR 有解析解,时间复杂度低.但是与 LRR 一样,得到的表示系数虽然保证了同一子空间的数据点之间的联系,但是不同子空间的数据点仍有较强的联系.Lu 等人[15]提出:聚类模型应保证同一子空间内数据点之间联系(如 LRR,LSR),同时又能像 SSC 一样,减小不同子空间数据点之间的联系.并且,通过引入正则项$\|XDiag(w)\|_*$解决这样的问题.李波等人[16]在低秩子空间聚类的基础上引入约束项 $tr(ZLZ^T)$,突出数据点间的局部特性,减小不同子空间数据点之间的联系,L 表示原数据集的拉普拉斯矩阵.Chen 等人[13]提出了 SLRR 算法,但本质还是求解二次规划的问题,只是将原数据进行低秩处理,不同子空间的数据点间仍有较强的联系.

1.2 LLC相关原理及方法

LLC 用在特征提取的编码阶段,对于每个数据点,LLC 寻求具有平滑的表示系数来表示此数据点的特征.

LLC 求解的优化问题为

$$\min_c \| x-Dc \|^2 +\frac{\lambda}{2}\| s\odot c \|^2 \quad \text{s.t. } \mathbf{1}^T c=1 \tag{2}$$

其中,x 表示数据点 x;c 表示数据点 x 的特征;D 表示字典,由训练得到;s 表示数据点 x 与字典 D 中元素的相似度,$s=\exp(dist(x,D)/\sigma)$,$dist(x,D)=[dist(x,d_1),\ldots,dist(x,d_j),\ldots,dist(x,d_M)]$,$dist(x,d_j)$表示数据点 x 与元素 d_j 的欧氏距离.

LLC 利用 D 中与 x 相近的 k 个元素作为字典求得 x 的表示系数,其他表示系数置为 0.对于相似的数据点,LLC 选取相似的近邻点作为字典,并且向量 s 对表示系数形成相似性约束.因此,对于相似的数据点会得到相似的特征.

2 GLSC 聚类算法

2.1 GLSC聚类模型

对于数据集 X,SLRR 求解的优化问题为

$$\min_Z \| X-XZ \|_F^2 +\frac{\lambda}{2}trace(Z^TZ) \quad \text{s.t. } X=XZ+E, Z=Z^T, rank(Z)\leqslant r \tag{3}$$

X表示原数据集,λ表示平衡因子.它有解析解 $Z^*=(A^TA+\lambda I)^{-1}A^TA$,其中,$A$ 是原数据集 X 的低秩形式.根据线性代数理论,若 A 是低秩的,则表示系数 Z 也是低秩的.A 可以通过 RPCA(robust principal component analysis)[17]或者 PCA(principal component analysis)[18]得到.对于每个数据点 x_i,公式(3)可以表示成

$$\min_Z \sum_{i=1}^N \| x_i-Xz_i \|_2^2 +\frac{\lambda}{2}\| z_i \|_2^2 \quad \text{s.t. } x_i=Xz_i+e_i, Z=Z^T, rank(Z)\leqslant r \tag{4}$$

SLRR 虽然利用ℓ_2范数寻求一个低秩对称的解,但是正如上文所述,它仍无法解决的是:z_i^* 虽然保证了 x_i 与同一子空间数据点的联系(z_i^* 是公式(4)的最优解,表示 x_i 的表示系数),但是与其他子空间的数据点仍有较强的联系,因此我们将正则项 $\| z_i \|_2^2$ 表示为 $\| w_i\odot z_i \|_2^2$.w_i 表示权值向量,⊙表示 w_i 与 z_i 中对应位置的元素两两相乘.若 w_i^k 表示 w_i 中的第 k 项,z_i^k 表示 z_i 中的第 k 项,则有:

$$\| w_i\odot z_i \|_2^2=\sum_{k=1}^d | w_i^k z_i^k |^2 \tag{5}$$

从公式(5)中可以看出:我们将表示系数 z_i 中第 k 项乘以权值 w_i^k,对于属于同一子空间的数据点,w_i^k 具有相似的值,使得他们具有相似的表示系数;对于不属于同一子空间的数据点,w_i^k 的值具有较大的差异性.这样求得的表示系数就会有较大的差异,即,弱化不同子空间数据点之间的联系.

因此,我们求解的优化问题为

$$\min_Z \sum_{i=1}^N \| x_i-Xz_i \|_2^2 +\lambda\| w_i\odot z_i \|_2^2 \quad \text{s.t. } \mathbf{1}^T z_i=1, \forall i, rank(Z)\leqslant r \tag{6}$$

其中,$X=[x_1\ x_2\ \ldots x_n]$,w_i 表示数据点 x_i 对应的权值向量,λ用来平衡两项,条件 $\mathbf{1}^Tz_i=1,\forall i$ 保证仿射空间不变性,$\mathbf{1}\in\mathbb{R}^{n\times 1}$ 表示 n 维全 1 的列向量.$rank(Z)\leqslant r$ 保证最终解的低秩性质,是通过寻求原数据集 X 的低秩形式保证的.它有解析解 $z_i=(A^TA+\lambda diag^2(w_i))A^Ta_i$, $z_i=\dfrac{z_i}{sum(z_i)}$.其中,$A$ 是原数据集 X 的低秩形式.为了有效求解仿射空间不变性,我们在得到 A 后,求解的优化问题变为

$$\left.\begin{aligned}&\min_{Z}\sum_{i=1}^{N}\| a_i - Az_i \|_2^2 + \lambda \| w_i \odot z_i \|_2^2 \\ &\text{s.t. } \mathbf{1}^T z_i = 1, \forall i\end{aligned}\right\} \tag{7}$$

利用拉格朗日方法,公式(7)可以写成

$$L = \| a_i \mathbf{1}^T z_i - Az_i \|_2^2 + \lambda \| w_i \odot z_i \|_2^2 + \theta(\mathbf{1}^T z_i - 1) \tag{8}$$

θ表示拉格朗日乘子,$A=[a_1\ a_2\ \ldots a_n]$是原数据集 X 的低秩形式,通过 RPCA 或者 PCA 求得.

将公式(8)对 z_i 求导,得:

$$\frac{\partial L}{\partial z_i} = 2D^{-1}z_i + \theta\mathbf{1} \tag{9}$$

其中,$D=((a_i\mathbf{1}^T-A)^T(a_i\mathbf{1}^T-A)+\lambda diag^2(w_i))^{-1}$.

令 $\frac{\partial L}{\partial z_i}=0$,得:

$$z_i = -\frac{1}{2}\theta D\mathbf{1} \tag{10}$$

公式(10)两边同时乘 $\mathbf{1}^T$,由于 $\mathbf{1}^Tz_i=1$,得到:

$$\theta = -\frac{2}{\mathbf{1}^T D\mathbf{1}} \tag{11}$$

将公式(11)带入公式(10),得:

$$z_i = \frac{D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}} \tag{12}$$

其中,$D=(C_i+\lambda diag^2(w_i))^{-1}$,$C_i=(a_i\mathbf{1}^T-A)^T(a_i\mathbf{1}^T-A)$.$C_i$ 表示数据点 a_i 的协方差矩阵,a_i 表示 A 中的第 i 列.

2.2 建立权值矩阵和相似矩阵

当前聚类算法中,很多学者对 ℓ_1 范数进行了加权方式的探讨,Xu 等人[19]利用 log-sum 启发式方法寻求对 ℓ_1 范数的加权方式,Liu 等人[20]通过 SIM(shape interaction matrix)[21]方法对 ℓ_1 范数加权.实验结果表明:对于 ℓ_1 范数,它们都是有效加权的方式.LLC 利用高斯核函数对 ℓ_2 范数加权,取得了更具平滑的表示系数,得到较好的分类结果.图 1 表示在 Extended YaleB 数据集前 10 个人脸对象上,使用高斯核函数建立权值矩阵时,参数σ对聚类准确率的影响.

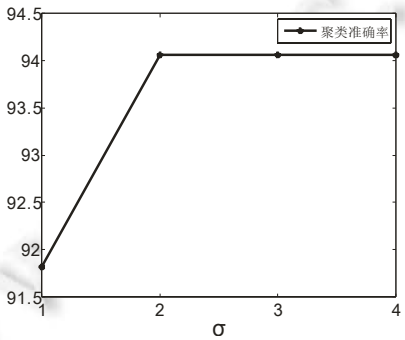


Fig.1 Influence of parameter σ for the accuracy of clustering, other two parameters n=5, λ=0.01

图 1 参数σ对聚类准确率的影响,算法中其他两个参数 n=5,λ=0.01

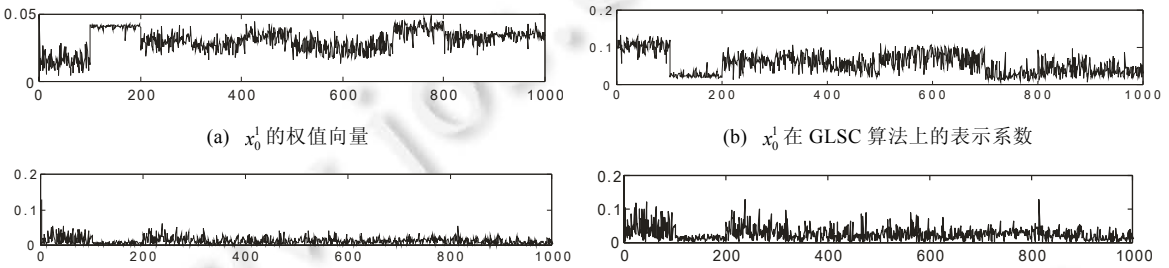
从图 1 中我们可以看出:当高斯核函数参数σ=2 时,聚类的准确率最高,为 94.06%.σ的值再往上增加时,聚类的准确率保持不变.经过实验发现:采用基于欧氏距离的加权方式时,聚类的准确率为 93.91%,两者相差不大,并且使用基于欧氏距离的加权方式时,算法只有两个参数,在其他数据集上也有类似的情况.因此,本文采用基于

欧氏距离的加权方式.假设 $X=[x_1\ x_2\ \ldots x_n]$表示 n 个数据点,$A=[a_1\ a_2\ \ldots a_n]$为其低秩形式,a_i 对应于数据点 x_i,因此其权值向量建立方式为:

(1) 计算 a_i 与 A 中每个数据点之间的欧氏距离,得到向量 $o_i\in\mathbb{R}^{n\times 1}$;

(2) 数据点 a_i 与其本身的距离设为∞,得到向量 o_i^*,令 x_i 对应的权值向量 $w_i=o_i^*$.

在稀疏子空间聚类中,作者为了避免平凡解,在优化问题中加入条件 $diag(Z)=0$[10].Peng 等人[22]为了避免平凡解,在 LSR 的基础上引入条件 $e_i^Tz_i=0$,$e_i\in\mathbb{R}^{n\times 1}$,$e_{ii}=1$,其余元素等于 0.在步骤(2)中,我们将数据点 x_i 与其本身的距离设为∞,这样,权值 w_i^i 对其本身的表示系数形成约束,从而避免平凡解.图 2 表示数据点 x_0^1(USPS 数据集上,数字 0 的第 1 张图片)的权值向量和各算法求得的表示系数.从图 2(a)中可以看出:数据点 x_0^1 与同一子空间中数据点的权值较小,与其他子空间的数据点有较大的权值.从图 2(b)中可以发现数据点 x_0^1 的权值向量对表示系数的约束作用:对于同一子空间的数据点有较小的权值,相应的表示系数比较大;对于不同子空间的数据点有较大的权值,相应的表示系数比较小,减小了不同子空间数据点之间的联系.



(a) x_0^1 的权值向量 (b) x_0^1 在 GLSC 算法上的表示系数

(c) x_0^1 在 LRR 算法上的表示系数 (d) x_0^1 在 LSR 算法上的表示系数

Fig.2

图 2

得到数据集 X 的系数矩阵 Z 后,本文采用文献[22]中建立相似矩阵的方式.保留 Z 中每列绝对值最大的 n 个实体,其他置为 0,得到 $\hat{Z}$,然后建立相似矩阵 J.

$$J=\frac{|\hat{Z}|+|\hat{Z}|^T}{2} \tag{13}$$

在谱聚类中,相似矩阵的高稀疏性有利于聚类的效果,建立矩阵 $\hat{Z}$ 用来保证相似矩阵 J 的稀疏性.谱聚类是一种基于图论的聚类方法,利用公式(13)建立对称矩阵 J 是基于其加权无向图的边权矩阵的对称性所决定的.$\hat{Z}$ 不是对称矩阵,不能保证数据点 i 与数据点 j 的相似度 $\hat{Z}_{ij}$ 等于数据点 j 与数据点 i 的相似度 $\hat{Z}_{ji}$.

整个算法流程见算法 1.

算法 1. GLSC.

输入:数据集 $X\in\mathbb{R}^{d\times n}$ 和聚类个数 K.N 表示数据点的个数;

1: 利用 RPCA 或者 PCA 求得原数据集 X 的低秩形式 A

2: for i=1 to N do

3: 求得 a_i 的权向量 w_i,

4: 利用公式(12)求解表示系数 z_i,

5: $Z=[Z\ z_i]$,

6: end for

7: 利用公式(13)建立相似矩阵 J

8: 将 J 用于谱聚类.

输出:数据集 X 的类分配.

3 组效应

本节证明 GLSC 具有组效应[12].组效应在子空间聚类中扮演重要的角色,对于相似的数据点,具有相似的表示系数.GLSC 的组效应特性可通过如下定理 1 得到.

定理 1. 给定一个数据样本 $y\in\mathbb{R}^d$、字典 $X\in\mathbb{R}^{d\times n}$ 和惩罚系数 λ.假设字典 X 中的每一列都正交化,z^* 是下面问题的最优解:

$$\min \| y-Xz \|_2^2+\lambda \| s\odot z \|_2^2 \tag{14}$$

则我们有:

$$\frac{\| z_i^*-z_j^* \|_2}{\| y \|_2}\leqslant\frac{1}{\lambda\min(s_i^2,s_j^2)}\sqrt{2(1-\cos(x_i,x_j))} \tag{15}$$

其中,s_i,s_j 表示权值向量 s 中第 i,j 项,即 z_i^*,z_j^* 对应的权值.

证明:

令 $L(z)=\| y-Xz \|_2^2+\lambda \| s\odot z \|_2^2$,因为 z^* 是问题(14)的最优解,所以 z^* 满足:

$$\frac{\partial L(z)}{\partial z}\Big|_{z=z^*}=0 \tag{16}$$

然后,我们有:

$$-2x_i^T(y-Xz^*)+2\lambda s_i^2z_i^*=0 \tag{17}$$

$$-2x_j^T(y-Xz^*)+2\lambda s_j^2z_j^*=0 \tag{18}$$

公式(17)与公式(18)相减,可得:

$$z_i^*-z_j^*\leqslant\frac{1}{\lambda\min(s_i^2,s_j^2)}(x_i^T-x_j^T)(y-Xz^*) \tag{19}$$

因为 X 的每一列都正交化,因此:

$$\| x_i^T-x_j^T \|_2=\sqrt{2(1-\cos(x_i,x_j))} \tag{20}$$

因为 z^* 是问题(14)的最优解,因此我们有:

$$\| y-Xz^* \|_2^2+\lambda \| s\odot z^* \|_2^2=L(z^*)\leqslant L(0)=\| y \|_2^2 \tag{21}$$

因此,$\|y-Xz^*\|_2\leqslant\|y\|_2$,然后通过公式(19)、公式(20)可以得到结论(15). □

4 实验结果与分析

本节通过实验验证 GLSC 聚类算法的有效性,第 4.1 节~第 4.3 节讨论各算法在 3 类数据集上的聚类效果.第 4.4 节讨论 GLSC 的参数选择.第 4.5 节讨论 GLSC 算法的时间复杂度.

本文使用聚类的准确率(accuracy)和 NMI(normalized mutual information)[23]评价算法的性能,在人脸数据库、手写数字数据库和运动分割上测试算法.其中,人脸数据库使用 Extend YaleB[24],AR[25]数据集.手写数字数据库使用 USPS[26],MNIST[27]数据集,运动分割使用 Hopkins 115[28]运动分割数据集.实验中的对比算法为 LRR,SSC,LSR,代码由原作者提供,且所有参数根据原论文中设置调到最优.另外,我们是利用 PCA 求得原数据集的低秩形式,在人脸数据集和手写数字集上将其映射到 $n\times6$ 维的低秩子空间中,n 表示聚类的类别数.在运动分割数据集上,将其映射到 12 维的低维子空间中.

表 1 给出了各算法的实验参数,需要注意的是:在 Extended YaleB 数据集和 AR 数据集上,SSC 和 LRR 在原论文中是对原数据集进行聚类;这里,我们将数据集降维到低维子空间后重新调整参数获取最优的结果.

Table 1 Parameters of all algorithms

表 1 各算法的参数设置

数据集	GLSC	SSC	LRR	LSR1	LSR2
YaleB(5)	n=5, λ=0.0001	α=100000	λ=2	λ=0.4	λ=0.4
YaleB(10)	n=5, λ=0.01	α=100000	λ=3	λ=0.004	λ=0.004
AR(10)	n=6, λ=0.00008	α=1000000	λ=2	λ=0.0007	λ=0.0004
AR(20)	n=6, λ=0.00001	α=1000000	λ=6	λ=0.0004	λ=0.006
MNIST	n=4, λ=0.4	α=6	λ=0.002	λ=15	λ=10
USPS	n=4, λ=0.05	α=1	λ=0.03	λ=130	λ=150
Hopkins115 (2 motions)	n=9, λ=0.0001	α=800	λ=4	λ=0.0046	λ=0.0048
Hopkins115 (3 motions)	n=10, λ=0.000001	α=800	λ=4	λ=0.0046	λ=0.0046

4.1 人脸数据集实验

本节在 YaleB 和 AR 人脸数据集上验证算法的有效性,其中,在 YaleB 数据库上,选取前 5 个和前 10 个人脸对象进行实验,分别包含 320 张和 640 张图片;AR 数据库中的有 50 位男性、50 位女性,每个人有 26 张图片,包括遮挡和戴眼镜的图片.为了建立与 YaleB 数据库上类似的子空间分割任务,在 AR 数据集上,我们选取前 10 个和前 20 个人脸对象实验.表 2 中的每条数据都是重复进行 10 次实验取平均得到.

Table 2 Accuracy and NMI of clustering on face databases (%)

表 2 不同算法在人脸数据集上聚类效果的比较 (%)

数据集	对象数	GLSC		SSC		LRR		LSR1		LSR2	
		AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI
YaleB	5	**98.75**	**96.24**	83.12	70.00	83.44	70.36	90.00	77.29	91.25	79.37
	10	**93.91**	**90.31**	65.94	57.51	61.56	54.58	68.80	59.91	72.19	64.40
AR	10	**93.46**	**92.31**	76.62	70.18	77.31	70.44	73.62	70.83	74.54	71.75
	20	**80.25**	**82.71**	71.79	73.62	70.44	75.55	65.69	74.13	71.03	71.56

从表 2 中可以看出:GLSC 在人脸数据集上,聚类的效果比当前流行的算法有较大的优势.从表中可以看出:在 YaleB 数据集上,当选取 5 个对象的人脸做实验时,GLSC 比 LSR2 有 7.5%的提升;但是当人脸对象提升到 10 个时,算法的优势明显,比 LSR2 有 21.72%的提升.在 AR 数据集上,当选取 10 个对象人脸实验时,GLSC 的聚类准确率比 LRR 有 16.15%的提升.当人脸对象提升到 20 个时,与 SSC 相比,有 8.46%的提升.综上所述,GLSC 算法在人脸数据集上有较好的聚类效果.

4.2 手写数字数据集实验

在手写数字数据集中,选取 USPS 数据集和 MNIST 数据集进行实验.它们都含有 10 个对象(0~9).由于它们样本数较多,所以在 USPS 数据集上,我们选取每个对象的前 100 张图片,共 1 000 张图片进行实验.在 MNIST 数据集上,我们选取每个对象的前 50 张图片,共 500 张图片进行实验.表 3 是各种算法在手写数字数据集上聚类效果的比较.

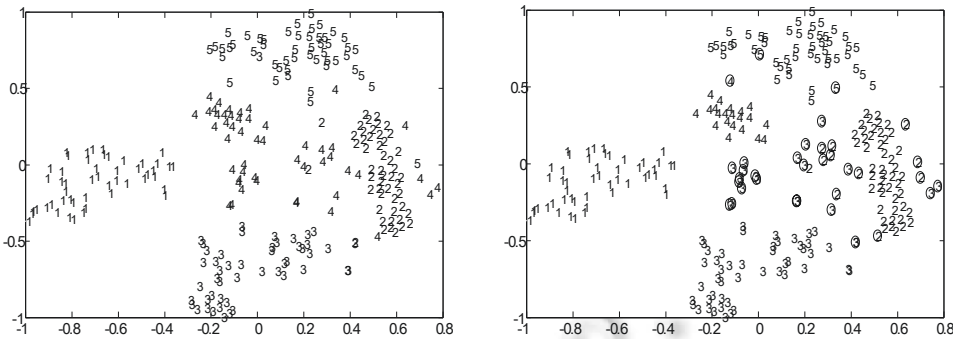
Table 3 Accuracy and NMI of clustering on handwrite databases (%)

表 3 不同算法在手写数字数据集上聚类效果的比较 (%)

数据集	GLSC		SSC		LRR		LSR1		LSR2	
	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI
MNIST	**72.80**	**66.69**	63.60	61.54	61.20	55.92	60.20	57.17	59.20	57.42
USPS	**83.00**	**82.63**	46.60	41.88	71.00	67.50	70.60	68.01	68.80	66.66

从表 3 中可以看出:所有的算法在手写数字数据集上都没有太好的聚类效果;但是与其他聚类算法相比,GLSC 算法仍有较好的聚类效果.在 MNIST 数据集上,GLSC 比 SSC 有 9.2%的提升.同样,在 USPS 数据集上,与 LRR 比较,在聚类的准确度上,GLSC 有 12%的提升.综上所述,GLSC 在手写数字数据集上有较好的聚类效果.

图 3 表示在 MNIST 数据集上,手写数字的可视化.图 3(a)表示 MNIST 数据集上,手写数字的真实分布.图 3(b)表示 GLSC 的聚类结果分布.图 3(b)中,带有圆圈的数字表示被错分的数字.



(a) MNIST 数据集的数据在低维空间的真实分布 (b) 在 MNIST 数据集上 GLSC 的聚类结果

Fig.3

图 3

4.3 运动分割数据集实验

在运动分割的实验中,采用聚类的错分率(1-accuracy)的均值和中值衡量各个算法的性能.Hopkins 155 数据集中包含 156 个运动序列,每个运动序列有 39 个~550 个数据点,其中包括 2 个运动序列和 3 个运动序列.我们使用在整个运动分割数据集上的平均错分率和中值衡量各个算法的有效性.表 4 是各个算法在运动分割数据集上聚类效果的比较.

Table 4 Segmentation errors of clustering on the Hopkins 155 database (%)

表 4 不同算法在运动分割数据上的聚类错误率 (%)

Algorithms	2 motions		3 motions	
	均值	中值	均值	中值
GLSC	2.43	**0**	**5.00**	**0.22**
SSC	**1.9**	**0**	5.10	1.09
LLR	3.64	0.22	9.43	3.7
LSR2	3.19	**0**	6.62	1.99

从表 4 中可以看出:在两个运动分割上,GLSC 的聚类错分率比 SSC 高 0.53%;但是在 3 个运动分割上,GLSC 比 SSC 有 0.1%的提升;并且在属性中值上,GLSC 只有 0.22%,具有一定的优越性.

4.4 算法参数的选择

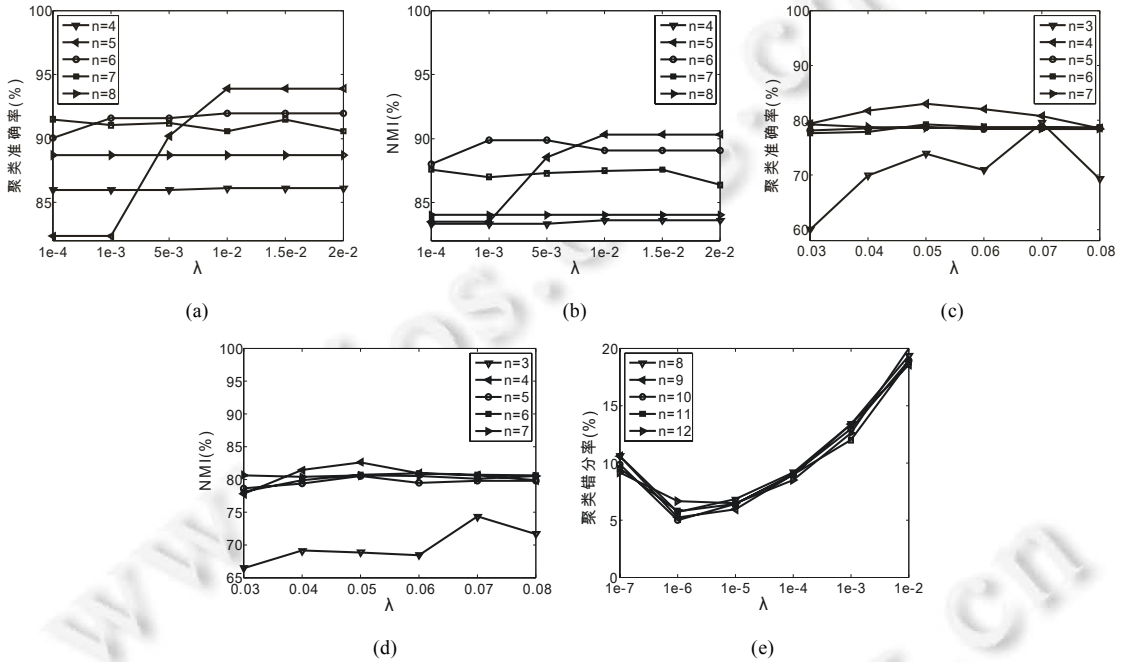
本节对 GLSC 算法在各个数据集上参数选择进行讨论.在人脸数据集上,选取 Extended YaleB 前 10 个人脸对象进行讨论.在手写数字数据集上,选取 USPS 数据集进行讨论.在运动分割数据集上,选取 3 个运动序列任务进行讨论.GLSC 算法在其他子空间分割任务上的参数选择不做讨论,参数选择方法与之类似.

图 4(a)表示在 ExtendedYaleB 数据集上,以 Accuracy 为评价指标时,算法参数λ和 n 对聚类性能的影响.从图中可以看出:当参数λ取值在 0.01~0.02 时,聚类的准确率比较稳定,对聚类性能的影响不大.同时可以看出,参数 n 对聚类效果的影响较大.当 n=5,λ=0.01 时,取得最好的聚类效果为 93.91%.从图 4(b)中可以看出:当λ取值在 0.01~0.02,n=5 时,NMI 也取得了最好的聚类效果,为 90.31%.在其他人脸分割任务上,采取同样的方法选取算法的参数,在 Extended YaleB 5 个人脸对象上,λ=0.0001,n=5.在 AR 数据集 10 个人脸对象上,λ=0.00008,n=6.20 个人脸对象上,λ=0.00001,n=6.

图 4(c)表示在 USPS 数据集上,以 Accuracy 为评价指标时,算法参数λ和 n 对聚类性能的影响.从图中可以发现:当 n=3 时,参数λ的取值对算法的稳定性的影响较大;但是当 n 取值在 4~7 时,参数λ的取值对聚类性能的影响较小.同时,当 n=4,λ=0.05 时,得到最好的聚类效果为 83%.同样,当以 NMI 为性能评价指标时,参数λ和 n 对聚类性能的影响类似.如图 4(d)所示:当 n=4,λ=0.05 时,取得最好的聚类性能为 82.63%.在 MNIST 数据集上,选取的参

数为 n=4,λ=0.4.

图 4(e)表示在 Hopkins155 3 个运动分割数据集上,以平均聚类错分率为性能评价指标时,算法参数λ和 n 对聚类性能的影响.当固定参数 n 时,参数λ对聚类效果的影响比较大;当λ从 1e−7 变化 1e−2 时,聚类的平均错分率从 9.87%下降到 5%,然后又逐渐提升到 18.96%.当固定参数λ时,参数 n 对聚类性能的影响不大.在 Hopkins155 2 个运动分割数据集上,参数的选取为 n=9,λ=0.0001.



(a) 在 Extended YaleB 前 10 个人脸数据集上,当以 Accuracy 为评价指标时,参数λ和 n 对聚类性能的影响
(b) 在 Extended YaleB 前 10 个人脸数据集上,当以 NMI 为评价指标时,参数λ和 n 对聚类性能的影响
(c) 在 USPS 数据集上,当以 Accuracy 为评价指标时,参数λ和 n 对聚类性能的影响
(d) 在 USPS 数据集上,当以 NMI 为评价指标时,参数λ和 n 对聚类性能的影响
(e) 在 Hopkins155 3 个运动数据集上,当以聚类的平均错分率为性能评价指标时,参数λ和 n 对聚类性能的影响

Fig.4 On different databases, the influence of two parameters on GLSC algorithm

图 4 在不同的数据集上,GLSC 算法中两个参数对聚类性能的影响

4.5 时间复杂度分析

为了验证各算法时间复杂度,在人脸数据集上,选取 Extended YaleB 前 10 个人脸对象进行实验.在手写数字数据集上,选取 USPS 数据集进行实验.在 Hopkins155 数据集上,选取 3 个运动分割进行实验.表 5 显示了各算法在不同数据集上的运行时间.

Table 5 Computing time of different algorithm on different databases (s)

表 5 各算法在不同数据集上的运行时间 (s)

算法	数据集		
	YaleB	USPS	Hopkins 155
GLSC	12.81	55.73	2.99
SSC	8.25	38.67	1.98
LRR	2.9	33.51	1.46
LSR	**0.03**	**0.15**	**0.02**

从表中可以看出,GLSC 的时间复杂度较高.公式(12)中每个数据点都需要计算协方差矩阵,并且求取协方

差矩阵后还要求逆矩阵,时间复杂度为 $o(n^3)$,n 表示数据点的个数.

因此,GLSC 在算法的运行时间上处于劣势.

5 结束语

本文利用 ℓ_2 范数,提出一种基于欧氏距离的且具有组效应的低秩子空间聚类算法,该算法在保证同一子空间数据点间联系的同时,减小不同子空间数据点之间的联系,并且在多个数据集上取得了较好的聚类效果.但是 GLSC 算法时间复杂度较高,这是由于对于每一个数据点,GLSC 要求解该数据点的协方差矩阵和求逆操作,因此在时间复杂度上,我们将对本算法做进一步优化.

References:

[1] Wang WW, Li XP, Feng XC, Wang SQ. A survey on sparse subspace clustering. ActaAutomatica Sinica, 2015,41(8):1373−1384 (in Chinese with English abstract). [doi: 10.16383/j.aas.2015.c140891]

[2] Vidal R, Ma Y, Sastry S. Generalized principal component analysis (GPCA). IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(12):1945−1959. [doi: 10.1109/TPAMI.2005.244]

[3] Rao S, Tron R, Vidal R, Ma Y. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(10):1832−1845. [doi: 10.1109/TPAMI.2009.191]

[4] Lu L, Vidal R. Combined central and subspace clustering for computer vision applications. In: Proc. of the 23rd Int'l Conf. on Machine Learning. New York: ACM Press, 2006. 593−600. [doi: 10.1145/1143844.1143919]

[5] Favaro P, Vidal R, Ravichandran A. A closed form solution to robust subspace estimation and clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 1801−1807. [doi: 10.1109/CVPR.2011.5995365]

[6] Elhamifar E, Vidal R. Clustering disjoint subspaces via sparse representation. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Dallas: IEEE, 2010. 1926−1929. [doi: 10.1109/ICASSP.2010.5495317]

[7] Yan J, Pollefeys M. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: Proc. of the European Conf. on Computer Vision. Berlin: Springer-Verlag, 2006. 94−106. [doi: 10.1007/11744085_8]

[8] Goh A, Vidal R. Segmenting motions of different types by unsupervised manifold clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007. 1−6. [doi: 10.1109/CVPR.2007.383235]

[9] Hu H, Lin Z, Feng J, Zhou J. Smooth representation clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 3834−3841. [doi: 10.1109/CVPR.2014.484]

[10] Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(11):2765−2781. [doi: 10.1109/TPAMI.2013.57]

[11] Liu G, Lin Z, Yan S, Sun J. Robust recovery of subspace structures by low-rank representation. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013,35(1):171−184. [doi: 10.1109/TPAMI.2012.88]

[12] Lu CY, Min H, Zhao ZQ, Zhu L, Huang DS, Yan S. Robust and efficient subspace segmentation via least squares regression. In: Proc. of the European Conf. on Computer Vision. Berlin: Springer-Verlag, 2012. 347−360. [doi: 10.1007/978-3-642-33786-4_26]

[13] Chen J, Zhang H, Mao H, Sang Y, Yi Z. Symmetric low-rank representation for subspace clustering. Neurocomputing, 2016,173:1192−1202. [doi: 10.1016/j.neucom.2015.08.077]

[14] Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y. Locality-Constrained linear coding for image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3360−3367. [doi: 10.1109/CVPR.2010.5540018]

[15] Lu C, Feng J, Lin Z, Yan S. Correlation adaptive subspace segmentation by trace lasso. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1345−1352. [doi: 10.1109/ICCV.2013.170]

[16] Li B, Lu CY, Leng CC, Jin LB. Robust low rank subspace clustering based on local graph Laplace constraint. Acta Automatica Sinica, 2015,41(11):1971−1980 (in Chinese with English abstract). [doi: 10.16383/j.aas.2015.c150031]

[17] Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? Journal of the ACM (JACM), 2011,58(3):11. [doi: 10.1145/1970392.1970395]

[18] Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2010,2(4):433−459. [doi: 10.1002/wics.101]

[19] Xu J, Xu K, Chen K, Ruan J. Reweighted sparse subspace clustering. Computer Vision and Image Understanding, 2015,138:25−37. [doi: 10.1016/j.cviu.2015.04.003]

[20] Liu B, Jing L, Yu J, Li J. Robust graph learning via constrained elastic-net regularization. Neurocomputing, 2016,171:299−312. [doi: 10.1016/j.neucom.2015.06.059]

[21] Costeira JP, Kanade T. A multibody factorization method for independently moving objects. Int'l Journal of Computer Vision, 1998,29(3):159−179. [doi: 10.1023/A:1008000628999]

[22] Peng X, Yi Z, Tang H. Robust subspace clustering via thresholding ridge regression. In: Proc. of the AAAI. Austin: AAAI, 2015. 3827−3833.

[23] Cai D, He X, Han J. Document clustering using locality preserving indexing. IEEE Trans. on Knowledge and Data Engineering, 2005,17(12):1624−1637. [doi: 10.1109/TKDE.2005.198]

[24] Georghiades AS, Belhumeur PN, Kriegman DJ. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(6):643−660. [doi: 10.1109/34.927464]

[25] Martinez AR, Benavente R. The AR face database, 1998. Technical Report, Computer Vision Center, 2007.

[26] Hull JJ. A database for handwritten text recognition research. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1994, 16(5):550−554. [doi: 10.1109/34.291440]

[27] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11):2278−2324. [doi: 10.1109/5.726791]

[28] Tron R, Vidal R. A benchmark for the comparison of 3-d motion segmentation algorithms. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007. 1−8. [doi: 10.1109/CVPR.2007.382974]

附中文参考文献:

[1] 王卫卫,李小平,冯象初,王斯琪.稀疏子空间聚类综述.自动化学报,2015,41(8):1373−1384. [doi: 10.16383/j.aas.2015.c140891]

[16] 李波,卢春园,冷成财,金连宝.基于局部图拉普拉斯约束的鲁棒低秩表示聚类方法.自动化学报,2015,41(11):1971−1980. [doi: 10.16383/j.aas.2015.c150031]

傅文进(1992−),男,江苏盐城人,硕士生,主要研究领域为聚类分析,人脸识别.

吴小俊(1967−),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为模式识别,计算机视觉,模糊系统,神经网络,智能系统.