计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500074

基于投影的鲁棒低秩子空间聚类算法

邢毓华　李明星
西安理工大学自动化与信息工程学院　西安 710048
(704189141 qq.com)

摘　要　随着大数据时代的来临,如何对海量高维数据进行有效的聚类分析并充分利用,已成为当下的热门研究课题。传统的聚类算法在处理高维数据时,聚类结果的精确度和稳定性较低,而子空间聚类算法通过分割原始数据的特征空间来得到不同的特征子集,可以大幅减小数据之间不相关特征对聚类结果的影响,挖掘出高维数据中不易展现的信息,在处理高维数据时具有显著的优势。针对现有基于图的子空间聚类算法在处理未知类型噪声以及复杂的凸问题时存在局限性的问题,在子空间聚类算法的基础上,结合空间投影理论,提出了一种基于投影的鲁棒低秩子空间聚类算法。首先对原始数据进行投影,利用编码消除投影空间的噪声,并对缺失的数据进行弥补;然后利用一种新的方法 l_2 图来构造稀疏相似图;最后在 l_2 图的基础上进行子空间聚类。该算法不需要对噪声的类型具有先验知识,且 l_2 图能够很好地描述高维数据稀疏性和空间分散的特征。选取 3 种人脸数据集作为实验数据集,首先确定影响聚类效果的最优参数,然后从准确度、鲁棒性、时间复杂度 3 个方面对算法进行验证。实验结果表明,在 3 种人脸数据集中混入未知类型的噪声时,该算法具有较高的准确率和较低的时间复杂度,并且具有好的鲁棒性。

关键词:高维数据;噪声;子空间聚类;空间投影;l_2 图
中图法分类号　TP311

Robust Low Rank Subspace Clustering Algorithm Based on Projection

XING Yu-hua and LI Ming-xing
College of Automation and Information Engineering,Xi'an University of Technology,Xi'an 710048,China

Abstract　With the advent of the era of big data,how to effectively cluster,analyze and effectively use massive amounts of high-dimensional data has become a hot research topic. When the traditional clustering algorithms are used to process high-dimensional data,the accuracy and stability of the clustering results are low. The subspace clustering algorithm can reduce the feature space of the original data to form different feature subsets,reduce the influence of uncorrelated features between data on clustering results. It can mine the information that is difficult to display in high-dimensional data,and has significant advantages in processing high-dimensional data. Aiming at the limitations of existing graph-based subspace clustering algorithms in dealing with unknown type noise and solving complex convex problems,based on subspace clustering algorithm,combined with spatial projection theory,this paper proposes a projection-based robust low-rank subspace clustering algorithm. Firstly,the original data is projected,the noise of the projection space is eliminated by coding and the missing data is compensated. Then a new method map is used to construct the sparse similarity l_2 graph,and finally the subspace clustering is performed on the basis of the l_2 graph. The algorithm does not need a priori knowledge of the type of noise,and the l_2 graph can well describe the characteristics of high-dimensional data sparsity and spatial dispersion. Three datasets of face recognition are selected as experimental datasets. Firstly,the optimal parameters affecting the clustering effect are determined,and then the algorithm is verified from three aspects:accuracy,robustness and time complexity. The experimental results show that the algorithm has high accuracy,low time complexity and good robustness,when the unknown type of noise is mixed in the datasets of face recognition.

Keywords　High dimensional data,Noise,Subspace clustering,Space projection,l_2 graph

1　引言

随着大数据技术的迅猛发展,各行业积累了海量的数据,其中高维数据占比迅速上升,遍及图像处理[1-2]、模式识别[3-4]以及人工智能[5]等多个领域。由于这类数据维数高,在分析挖掘过程中存在较大困难,如何对这些高维数据进行挖掘分

到稿日期:2019-05-15　返修日期:2019-09-03　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(51307140)
This work was supported by the National Natural Science Foundation of China (51307140).
通信作者:李明星(15129012577@163.com)

析从而进行有效利用成为当下的研究热点。子空间聚类以低维空间为基础,对高维数据进行集群分割,使每个集群与一个线性独立的子空间相对应,只在相关的子空间上进行挖掘任务,可对高维数据进行有效聚类。

现存的关于子空间聚类算法的研究从子空间聚类机制和应用层面出发,主要可分为 4 种:矩阵分解方法[6]、迭代方法[7]、统计方法[8-9]和谱聚类方法[10]。矩阵分解方法只适用于数据相互独立的子空间,有一定的局限性。若假定的子空间内的数据及噪声分布不清晰,则迭代方法无法得到最优解。统计方法由于模型限制不能被应用于大规模数据。谱聚类方法将数据和图相结合,通过图谱划分实现最优聚类,已成为当下的研究热点。

一般以谱聚类为基础的子空间聚类算法有两种典型的表示方式:1)基于低秩模型的低秩表示(Low Rank Representation,LRR)[11-12];2)基于稀疏模型的稀疏表示(Sparse Representation,SR)[13]。该算法的核心是构建亲和矩阵,常用的方法有两种:1)利用欧氏距离的局部距离法,如 Laplace 特征图、k 临近等;2)基于线性表示的全局线性表示法[14],如低秩子空间聚类(Low-rank Subspace Clustering,LSC)[15]表示、稀疏子空间聚类(Sparse Subspace Clustering,SSC)表示。低秩子空间聚类主要根据奇异值的稀疏性来获取所需的子空间结构,筛选出该组数据的最低秩。稀疏子空间聚类通过建立低维子空间的映射模型,找出其分布规律,求得映射系数,来构建亲和矩阵,从而构建相似图,并对图进行分割。若将每个数据点都表示为其他数据点的线性组合,则子空间相互独立或者不相交并且不含有噪声时,SSC 和 LRR 均可达到很好的聚类效果。但在实际应用中,数据集还包含各类噪声,使数据同时位于两个或多个子集,导致不同集群的数据边缘权重变高,在很大程度上降低了算法的效率。本文提出了基于投影的低秩子空间聚类算法,其结合新的构图方式,采用编码方式去除投影空间的噪声,同时对缺失数据进行弥补,且不需要噪声类型的先验知识,因而具有较好的聚类效果。

2 相关工作

基于图的谱聚类算法应用广泛,其主要分为 3 类:聚类分析[16]、图形追踪[17]和人脸识别[18]。这类算法的核心在于构造相似图,在实际应用中图对数据集的表示度决定了算法的优劣,构建方法主要有局部距离法和全局线性表示法。对于局部距离法而言,其衡量标准如下:

$$A_{ij}=\exp(-dist_{ij}^2) \tag{1}$$

其中,A_{ij} 代表两数据点 i 和 j 的关联度。$dist_{ij}$ 为某一距离求解公式,如欧氏距离公式等。该方法可用于构造 Laplace 特征图、k 临近的相似图。然而,因局部距离法对噪声和离群值的敏感性,导致其无法处理被噪声损坏的数据。

全局线性表示法通过采用不同范数对其表示矩阵进行正则化来构建相似图,其相似性是数据自适应的,该特性可以增强算法的鲁棒性。全局线性表示法已被广泛用于高维数据聚类分析,其具体表示为:

$$\min_{Z}\alpha\|\boldsymbol{X}-\boldsymbol{A}(\boldsymbol{x})\|_p+\boldsymbol{B}(\boldsymbol{x},\boldsymbol{z})$$
$$\text{s.t. } \boldsymbol{Z}\in\boldsymbol{C} \tag{2}$$

其中,$\boldsymbol{X}$ 代表数据矩阵,$\boldsymbol{A}(\boldsymbol{x})$ 表示字典,一般等于 $\boldsymbol{X}$;$\|\cdot\|_P$ 代表某种范数;$\boldsymbol{B}(\boldsymbol{x},\boldsymbol{z})$ 为正则项;$\boldsymbol{C}$ 为约束集合;α 为惩罚因子。全局线性表示法的典型模型包括局部线性表示(Lolocality Linear Representation,LLR)[19]、稀疏子空间聚类(SSC)和低秩表示(LRR)等。

LLR 在特征提取阶段可充分利用图像的几何结构信息,打破了全局表示的局限性,其具体表示如下:

$$\min_{c}\|\boldsymbol{a}-\boldsymbol{D}\boldsymbol{c}\|_2^2$$
$$s.t.\ \boldsymbol{c}^{\mathrm{T}}1=1 \tag{3}$$

其中,$\boldsymbol{a}$ 表示数据点,$\boldsymbol{c}$ 表示数据点的特征,$\boldsymbol{D}$ 表示由训练得到的字典矩阵。

SSC 的基本思想是将原数据集的样本用其他样本的线性组合来表示,其已被广泛应用于子空间聚类相似图的构建和子空间学习[20],具体公式如下:

$$\min_{\boldsymbol{Z},\boldsymbol{E}}\|\boldsymbol{Z}\|_1+\lambda\|\boldsymbol{E}\|_1$$
$$\text{s.t. } \boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z},\boldsymbol{Z}_{ii}=0 \tag{4}$$

其中,$\boldsymbol{X}$ 代表输入数据;$\boldsymbol{Z}$ 代表子空间表示矩阵,且 $\boldsymbol{Z}$ 具有对角结构,$\boldsymbol{Z}_{ii}$ 为 0 表示每个数据只能由自己表示;$\boldsymbol{E}$ 为噪声或奇异样本。

LRR 通过低秩约束条件对整个数据集的全局结构约束进行表示,使得所求子空间的表示矩阵 $\boldsymbol{Z}$ 的秩最小,且根据秩的最小性质,可用核范数来进行矩阵秩的最小化替换,这是一种更简便且有效的方法。该方法的具体公式如下:

$$\min_{\boldsymbol{Z},\boldsymbol{E}}\boldsymbol{J}(\boldsymbol{Z},\boldsymbol{E})=\|\boldsymbol{Z}\|_*+\lambda\|\boldsymbol{E}\|_p$$
$$\text{s.t. } \boldsymbol{A}=\boldsymbol{A}\boldsymbol{Z}+\boldsymbol{E},\boldsymbol{Z}^{\mathrm{T}}1=1 \tag{5}$$

其中,矩阵 $\boldsymbol{A}$ 是给定的数据集;矩阵 $\boldsymbol{Z}$ 是矩阵 $\boldsymbol{A}$ 的一个低秩表示;$\|\boldsymbol{Z}\|_*$ 是矩阵 $\boldsymbol{Z}$ 的核范数,其值代表矩阵 $\boldsymbol{Z}$ 的奇异值之和;$\|E\|_p$ 表示数据中的噪声和误差。

从上文可以发现,SSC 和 LRR 都是从输入空间的角度出发进行建模的,以弥补数据缺失和数据污染。本文从投影空间这个新角度出发,利用空间投影对数据集的字典矩阵进行预处理,并对原始数据进行压缩,从而将大量噪声和奇异值进行过滤,得到一个降维的近似原始数据,以弥补数据缺失和去除污染数据,从而得到更好的聚类效果。

3 基于投影的鲁棒低秩子空间聚类算法

本文提出了基于投影的鲁棒低秩子空间聚类算法(Robust Low Rank Subspace Clustering Based on Projection,RLRSCP),其利用空间投影对数据进行预处理以得到降维的优化数据。为解决空间投影对外点过于敏感的问题,因 l_2 范数为可解析求解范数,所以所提算法在 l_2 范数的基础上进行投影,这样可以在降低维数的同时降低对外点的敏感度,提高算法的鲁棒性,使子空间表示的亲和矩阵拥有更为精准的低秩结构。在此基础上构建相似图,简称为 l_2 图,从而得到对应的拉普拉斯矩阵 $\boldsymbol{L}$,通过对 $\boldsymbol{L}$ 的前 n 个特征向量进行 k-subspace 聚类,得到最终聚类结果。

3.1 投影子空间理论

本文考虑到总子空间 $\boldsymbol{S}_D$ 中的非零数据点 $\boldsymbol{a}$ 的分布情况,将 $\boldsymbol{a}$ 所在子空间记为 $\boldsymbol{S}_{D_a}$,其表示矩阵为 $\boldsymbol{D}_a$;$\boldsymbol{S}_D$ 中除 $\boldsymbol{S}_{D_a}$ 以外的子空间记为 $\boldsymbol{S}_{D_{-a}}$,其表示矩阵为 $\boldsymbol{D}_{-a}$,即 $\boldsymbol{S}_D=[\boldsymbol{S}_{D_a},\boldsymbol{S}_{D_{-a}}]$。

那么对于子空间中的 $\boldsymbol{a}$ 而言，存在的优化问题如下：

$$\begin{bmatrix} \boldsymbol{c}^*_{D_a} \\ \boldsymbol{c}^*_{D_{-a}} \end{bmatrix} = \arg\min \left\| \begin{bmatrix} \boldsymbol{c}_{D_a} \\ \boldsymbol{c}_{D_{-a}} \end{bmatrix} \right\|_p$$

即：

$$\min \| \boldsymbol{c} \|_p \quad \text{s.t } \boldsymbol{a} = \boldsymbol{D}\boldsymbol{c} \tag{6}$$

其最优解为 $\boldsymbol{c}^* = \begin{bmatrix} \boldsymbol{c}^*_{D_a} \\ \boldsymbol{c}^*_{D_{-a}} \end{bmatrix}$，满足 $[\boldsymbol{c}^*_{D_a}]_{r_a,1} > [\boldsymbol{c}^*_{D_{-a}}]_{1,1}$。其中 $[\boldsymbol{c}^*_{D_a}]_{-r_a,1}$ 代表 $\boldsymbol{c}^*_{D_a}$ 中前 r_a 个绝对值最大的项。

本节通过对数据分布情况的整体考虑，将数据点 a 的位置分为以下两种情况：

(1)数据点 $\boldsymbol{a}$ 位于子空间 S_{D_a} 与 $S_{D_{-a}}$ 的交集，即 $\boldsymbol{a} \in \{\boldsymbol{S} | \boldsymbol{S} = \boldsymbol{S}_{D_a} \cap \boldsymbol{S}_{D_{-a}}\}$。

(2)数据点 $\boldsymbol{a}$ 仅位于子空间 S_{D_a} 或 $S_{D_{-a}}$ 内，即 $\boldsymbol{a} \in \{\boldsymbol{S} | \boldsymbol{S} = \boldsymbol{S}_{D_a} \backslash \boldsymbol{S}_{D_{-a}}\}$。

下面对最优解进行证明：

$$\begin{aligned} \boldsymbol{a} &= DC^* = [D_a \quad D_{-a}] \begin{bmatrix} \boldsymbol{C}^*_{D_a} \\ \boldsymbol{C}^*_{D_{-a}} \end{bmatrix} \\ &= D_a C^*_{D_a} + D_{-a} C^*_{D_{-a}} \end{aligned} \tag{7}$$

由式(7)得：

$$\boldsymbol{a} - \boldsymbol{D}_{-a}\boldsymbol{C}^*_{D_{-a}} = \boldsymbol{D}_a \boldsymbol{C}^*_{D_a} \tag{8}$$

令

$$y = a - D_{-a} C^*_{D_{-a}}, y \in S_{D_a} \tag{9}$$

由式(8)可得：

$$\bar{y} = D_{-a} C^*_{D_{-a}}, \bar{y} \in S_{D_{-a}} \tag{10}$$

y 与 $\bar{y}$ 存在如下最优解：

$$y^* = D_a X_{D_a}, \bar{y}^* = D_{-a} X_{D_{-a}} \tag{11}$$

其中，X_{D_a} 和 $X_{D_{-a}}$ 表示 y 和 $\bar{y}$ 在式(6)中的解。

将 y^* 代入式(9)可得：

$$a = [D_a \quad D_{-a}] \begin{bmatrix} C^*_{D_a} + X_{D_a} \\ 0 \end{bmatrix} \tag{12}$$

将 $\bar{y}^*$ 代入式(10)可得：

$$a = [D_a \quad D_{-a}] \begin{bmatrix} C^*_{D_a} \\ X_{D_{-a}} \end{bmatrix} \tag{13}$$

其中，$\begin{bmatrix} C^*_{D_a} + X_{D_a} \\ 0 \end{bmatrix}, \begin{bmatrix} C^*_{D_a} \\ X_{D_{-a}} \end{bmatrix}$ 为式(6)的两个合理的解。

根据范数不等式原则：

$$\begin{aligned} \left\| \begin{bmatrix} C^*_{D_a} + X_{D_a} \\ 0 \end{bmatrix} \right\|_2 &\leqslant \| C^*_{D_a} \|_2 + \| X_{D_a} \|_2 \\ &\leqslant \| C^*_{D_a} \|_2 + \| X_{D_{-a}} \|_2 < \| C^*_{D_a} \|_2 + \| C^*_{D_{-a}} \|_2 \\ &= \left\| \begin{bmatrix} C^*_{D_a} \\ C^*_{D_{-a}} \end{bmatrix} \right\| \end{aligned} \tag{14}$$

由于涉及数据点的分布情况比较复杂，考虑到其与投影空间存在契合度的问题，给出如下两个定义。

定义 1 子空间 S_i 与 S_j 的最小夹角为：

$$\theta_{\min} = \min_{V_i \in S_i, V_j \in S_j} \{\arccos(\frac{V_i^{\mathrm{T}} V_j}{\| V_i \|_2 \| V_j \|_2})\} \tag{15}$$

其中，S_i 的维数为 r_1，S_j 的维数为 r_2，且 $r_1 \leqslant r_2$。

定义 2 矩阵 D_a 的奇异值分解为 $D_a = U\Sigma_{r_a} V^{\mathrm{T}}$，$D_a$ 的秩是 r_a，$\Sigma_{r_a} = diag(\sigma_1, \sigma_2, \cdots \sigma_{r_a})$，$\sigma$ 表示 D_a 的奇异值。$\min \sigma(D_a) \geqslant \cos\theta_{\min} \max \| D_{-a} \|_2$ 时，$[\boldsymbol{c}^*_{D_a}]_{r_a,1} > [\boldsymbol{c}^*_{D_{-a}}]_{1,1}$ 成立，$\| D_{-a} \|_2$ 表示 D_{-a} 中列向量的 ℓ_2 范数。具体证明如下：

对于情况(1)有：

$$a = D_a X_{D_a} = U\Sigma_{r_a} V^{\mathrm{T}} X_{D_a} \tag{16}$$

由式(16)得：

$$X_{D_a} = V\Sigma^{-1}_{r_a} U^{\mathrm{T}} a \tag{17}$$

由范数的性质得：

$$\begin{aligned} \| X_{D_a} \|_p &\leqslant \| X_{D_a} \|_1 \leqslant \sqrt{n} \| X_{D_a} \|_2 \leqslant \sqrt{n} \| V\Sigma^{-1}_{r_a} U^{\mathrm{T}} a \|_2 \\ &\leqslant \frac{\sqrt{n}}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_{r_a}^2}} \| a \|_2 \\ &\leqslant \sigma^{-1}_{\min}(D_a) \| a \|_2 \end{aligned} \tag{18}$$

其中，$\sigma_{\min}(D_a) = \sigma_{r_a}(D_a)$ 是 D_a 最小的非零奇异变量。

对于情况(2)有：

$$\begin{aligned} &a = D_{-a} X_{D_{-a}} \\ &\| a \|_2^2 = a^{\mathrm{T}} D_{-a} X_{D_{-a}} \leqslant \| D_a^{\mathrm{T}} a \|_\infty \| X_{D_a} \|_1 \\ &\| D_a^{\mathrm{T}} a \|_\infty = \max(|[D_a]_1^{\mathrm{T}} a|, |[D_a]_1^{\mathrm{T}} a|, \cdots) \end{aligned} \tag{19}$$

$$\leqslant \cos\theta_{\min} \| D_a \|_{\max,2} \| a \|_2 \tag{20}$$

通过以上的理论描述及证明可以发现，基于 ℓ_2 范数的空间投影得到的投影系数恒大于噪声的投影系数，因此对于较小的系数，本文直接将其编码为 0，以此从投影空间消除噪声，这样的处理方式不需要提前知道噪声类型，有极大的便捷性。

3.2 算法描述

假设 $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n\}$ 是样本数据点构成的集合，其位于总子空间 S_D 中。令 $\boldsymbol{A}_i = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{i-1}, 0, \boldsymbol{a}_{i+1}, \cdots, \boldsymbol{a}_n\}$ 为数据点 a_i 的字典矩阵，则所求目标函数可表示为：

$$\begin{aligned} &\min_{c_i} \| a_i - A_i c_i \|_2^2 + \frac{1}{2}\lambda \| \boldsymbol{c}_i \|_2^2 \\ &\text{s.t. } \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{c}_i = 0 \end{aligned} \tag{21}$$

其中，$\boldsymbol{e}_i \in \boldsymbol{R}^{n \times 1}$，$\boldsymbol{e}_{ii} = 1$，其余元素为 0。

利用拉格朗日乘子法，可得：

$$\mathrm{L}(\boldsymbol{c}_i) = \| \boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{c}_i \|_2^2 + \lambda \| \boldsymbol{c}_i \|_2^2 + l\boldsymbol{e}_i^{\mathrm{T}} c_i \tag{22}$$

对其关于 $\boldsymbol{c}_i$ 求偏导，令 $\frac{\partial L(\boldsymbol{c}_i)}{\partial(\boldsymbol{c}_i)} = 0$，可得：

$$(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \lambda \boldsymbol{E})\boldsymbol{c}_i - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i + \frac{1}{2} l \boldsymbol{e}_i^{\mathrm{T}} = 0 \tag{23}$$

由式(23)可得：

$$\boldsymbol{c}_i = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \lambda \boldsymbol{E})^{-1} (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i - \frac{1}{2} l \boldsymbol{e}_i) \tag{24}$$

其中，$\boldsymbol{E} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$，$\boldsymbol{E} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$，$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为矩阵 $\boldsymbol{A}$ 奇异值分解的两个正交矩阵。

由式(21)可得：

$$l = \frac{2\boldsymbol{e}_i^{\mathrm{T}} (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i + \lambda \boldsymbol{E})^{-1} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i}{\boldsymbol{e}_i^{\mathrm{T}} (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i + \lambda \boldsymbol{E})^{-1} \boldsymbol{e}_i} \tag{25}$$

令 $\boldsymbol{P} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \lambda \boldsymbol{E})^{-1}$，$\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{A}^{\mathrm{T}}$。由式(24)、式(25)可得：

$$\boldsymbol{c}^* = \boldsymbol{P}[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{a}_i - \frac{\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{a}_i\boldsymbol{e}_i}{\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{P}\boldsymbol{e}_i}] \tag{26}$$

通过式(26)可求得每一个点 $\boldsymbol{a}_i$ 对应的 $\boldsymbol{c}_i$，对其进行投影，令：

$$\boldsymbol{c}_i = \boldsymbol{\Psi}_{\eta,\gamma}(\boldsymbol{c}_i) \tag{27}$$

映射 $\boldsymbol{\Psi}_k(\boldsymbol{c}_i)$对 $\boldsymbol{c}_i$ 进行如下处理:

(1)设定阈值参数 η,对其中前 η 项大的元素保持原有值;

(2)对其余元素设定阈值参数 γ,对 $\boldsymbol{c}_i$ 中大于$\frac{\hat{c}}{\gamma}$的元素进行保留,将其他元素设置为 0,其中 $\eta>0$ 且取值必须为正整数,$\gamma\geqslant 1.0$ 且 $\gamma\in R$,$\hat{c}$为 c_i 中最大元素的值。

基于投影的低秩子空间聚类算法如算法 1 所示。

算法 1 基于投影的低秩子空间聚类算法

输入:总子空间 $S=\{s_i\}_{i=1}^n$ 中的数据点的集合 $A=\{a_i\}_{i=1}^n$,设定参数 λ,η,γ

输出:聚类结果

1. 设定权衡参数 λ,以及阈值参数 η 和 γ。
2. 根据式(26)计算样本数据点 $\mathbf{a}_i$ 对应的 c_i。
3. 利用式(27)对 $\mathbf{c}_i$ 进行编码,消除相关噪声和奇异值的影响。
4. 根据亲和矩阵 $\mathbf{W}_{ij}=\frac{1}{2}[|\mathbf{c}_{ij}|+|\mathbf{c}_{ji}|]$构建 ℓ_2 图。
5. 构建拉普拉斯矩阵 $\mathbf{L}=\Sigma^{-\frac{1}{2}}\mathbf{W}\Sigma^{-\frac{1}{2}}$,其中 $\Sigma=\mathrm{diag}\{\sum_{i=1}^{n}\mathbf{W}\}$。
6. 计算拉普拉斯矩阵 $\mathbf{L}$ 的特征向量,并对其所有项中的前 k 个特征值对应的特征向量构成的矩阵进行 k-subspace 聚类,最终得到所有的聚类结果。

4 实验验证

本节通过实验对本文提出的基于投影的鲁棒低秩子空间聚类算法进行验证。实验选用子空间聚类算法常用的 ORL、Extended Yale Database B 以及 AR 等 3 个人脸数据集作为实验数据集。实验首先利用 ORL 人脸数据集确定影响本文算法聚类效果的最优参数,然后从聚类准确度、算法鲁棒性和算法时间复杂度 3 个方面将所提方法与其他新近提出的子空间聚类算法 LRR,SSC 和 LSR 进行比较,从而验证本文算法的正确性和有效性。实验中,鲁棒性是通过噪声添加损坏样本来进行实验评估的,这里考虑高斯噪声和随机噪声两种噪声。

(1)参数设置:为保证实验结果的公平性,在将本文算法与经典子空间聚类算法进行比较时,依据原有论文,将 LRR,SSC,LSR 3 种算法的参数均设置到最优。

(2)性能评价:本文依据子空间聚类算法的特性,采用聚类算法常用的两个指标,即准确性(Cluster Accuracy,CA)和标准化互信息 NMI(Normalized Mutual Information),来评价算法的性能。这两个指标的优势是可以依据现有的数据对聚类效果进行直接评价,更贴近聚类算法的实际表现。

(3)参数确定:ℓ_2 图建立的 3 个关键参数为 λ,η 和 γ,其中 λ 为权衡参数,η 和 γ 为阈值参数。一般来说,λ 参数与数据中的噪声成正比,数据中的噪声越大,λ 参数的取值就越大。对于阈值参数 η 和 γ,其取值原则为使向量 c_i 中的非零项趋近于相应子空间的维数。

(4)鲁棒性:本文实验通过在数据集中添加高斯噪声和随机噪声来破坏样本,从而对聚类算法的鲁棒性进行评估。

4.1 获取最优参数

为了确定参数 λ,η 和 γ 对本文算法聚类效果的影响,从而确定其最优参数,从 ORL 人脸数据集中选取 20 个人脸对象进行实验。

图 1 给出了固定参数 $\eta=4,\gamma=4$ 时,参数 λ 值对聚类准确性的影响。可以看出,当 λ 取值为 0.01～0.7 时,聚类准确度保持在较高水平;当 $\lambda=0.3$ 时聚类准确性达到最高。

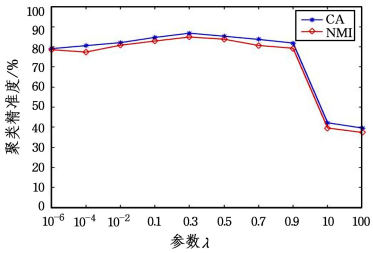


图 1 参数 λ 对聚类准确性的影响

Fig. 1 Influence of parameter λ on clustering precision

图 2 给出了当固定参数 $\lambda=0.3$ 时,在不考虑 γ 取值的情况下,参数 η 值对聚类准确性的影响。可以看出,当 η 取值为 6～10 时,聚类准确度保持在较高水平;当 $\eta=8$ 时聚类准确性达到最高。

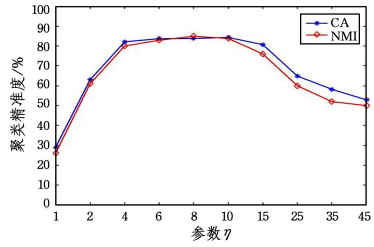


图 2 参数 η 对聚类准确性的影响

Fig. 2 Influence of parameter η on clustering precision

阈值参数 η 和 γ 都可以提高模型的鲁棒性,可以除去 c_i 中较小的对应噪声的元素。但是,较大的 η 和 γ 会影响模型的识别能力,而较小的 η 和 γ 又不能提供足够的代表性。因此,对参数 η 和 γ 同时设置对应的阈值比单独设置其中一个参数的阈值更有优势:如果只设置阈值参数 η,那么每个点所对应的子空间的维数都相同且是确定的值,这完全不符合实际情况;如果只设置阈值参数 γ,则存在 c_i 的最大项远远大于其他项的情况,不具有足够的代表性。

图 3 给出了当固定参数 $\lambda=0.3$ 时,在不考虑参数 η 取值的情况下,参数 γ 值对聚类准确性的影响。可以看出,当 γ 取值为2～6 时,聚类准确度保持在较高水平;当 $\gamma=4$ 时聚类准确性达到最高。

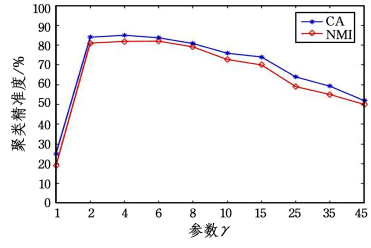


图 3 参数 γ 对聚类准确性的影响

Fig. 3 Influence of parameter γ on clustering precision

对于阈值参数 η 和 γ,由于其中一个参数的取值会对另一个参数造成影响,为了对两个参数进行调优使得聚类效果最好,本文通过固定一个参数来研究另一个参数变化对聚类

效果的影响，不断重复此步骤，直到取得最优参数。图 4 给出了固定参数 η 为 8 时，参数 γ 对聚类准确性的影响。

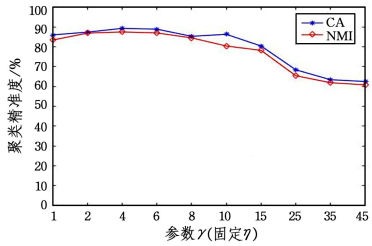


图 4 参数 γ 对聚类精准性的影响($\eta=8$)

Fig. 4 Influence of parameter γ on clustering precision

从图 4 可以看出，当固定参数 $\eta=8$ 且 γ 取值为 2～6 时，聚类准确度保持在较高水平。

4.2 算法性能验证

本节对所提出的基于投影的鲁棒低秩子空间聚类算法的性能进行验证。选取人脸数据集 AR 中 80 个人的人脸数据，每人选取 20 张不同的照片，通过不同样本数下的聚类结果和模型的 *NMI* 值来对算法的性能进行评估，实验结果如图 5 和图 6 所示。

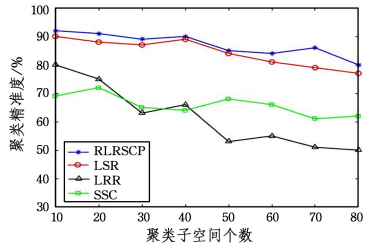


图 5 各个聚类模型在不同样本数下的聚类结果

Fig. 5 Clustering results of each clustering model under different sample numbers

从图 5 可以看出，随着聚类子空间数目的增大，几种算法的聚类准确性基本呈逐步降低的趋势，但是本文算法的准确性高于其他 3 种算法，其聚类性能更好。

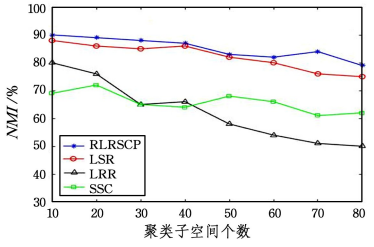


图 6 各个聚类模型在不同聚类子空间个数下的 *NMI* 值

Fig. 6 NMI value of each cluster in the number of different cluster subspace

从图 6 可以看出，本文算法的 *NMI* 值随着聚类子空间数目的增大有所降低，但变化幅度较小，表明该算法具有较好的鲁棒性。

4.3 鲁棒性验证

本节对所提出的基于投影的鲁棒低秩子空间聚类算法的鲁棒性进行验证。选取 Extended Yale Database 人脸数据集中 30 个人的人脸数据，并从每人的照片中选取 30 幅进行噪声污染，其中噪声分为高斯白噪声和随机噪声。通过算法的 *AC* 值和 *NMI* 值来对算法的鲁棒性进行评估，实验结果如表 1 和表 2 所列。

从表 1 和表 2 可以看出，各种聚类算法在受高斯噪声影响时的聚类准确性和 *NMI* 均优于受随机噪声影响的聚类结果。不论是在高斯噪声的影响下还是在随机噪声的影响下，本文算法的聚类效果均优于另外几种算法，且聚类效果的浮动幅度较小，表明其具有较好的鲁棒性。

表 1 各种算法聚类的 *AC* 值

Table 1 AC value of clustering results

(单位：%)

噪声	ρ	PLPSCP	LSR	LRR	SSC
高斯噪声	10	89.15	80.13	86.23	71.43
	20	88.62	73.14	85.18	62.82
	30	84.31	65.48	82.60	57.41
	40	79.85	60.72	77.15	53.14
随机噪声	10	83.14	73.45	80.46	64.12
	20	70.13	58.85	68.45	55.14
	30	50.45	48.23	48.45	39.95
	40	39.94	30.23	32.87	25.46

表 2 各种算法聚类的 *NMI* 值

Table 2 *NMI* value of clustering results

(单位：%)

噪声	ρ	PLPSCP	LSR	LRR	SSC
高斯噪声	10	84.82	75.31	80.69	67.81
	20	82.34	68.64	79.83	61.29
	30	77.68	60.97	75.27	58.16
	40	73.91	56.13	70.82	51.33
随机噪声	10	75.73	69.88	72.38	61.84
	20	68.59	56.69	65.15	50.69
	30	62.40	47.22	52.83	48.26
	40	58.65	42.71	44.76	41.93

4.4 时间复杂度验证

本节对所提出的基于投影的鲁棒低秩子空间聚类算法的时间复杂度进行验证。本文从 ORL 人脸数据集和 AR 人脸数据集中各选取 20 个人的人脸数据进行实验，结果如表 3 所列。

表 3 各种算法的运行时间

Table 3 Running time of each algorithm

(单位：s)

算法 \ 数据集	AR	ORL
RLRSCP	53.20	37.86
LSR2	60.21	51.23
LRR	78.32	64.39
SSC	92.86	81.73

从表 3 可以看出，本文算法不管是在 AR 人脸数据集上还是在 ORL 人脸数据集上，运行速度都高于其他几种算法，表明本文算法具有较高的时间效率。

结束语 现有的大多数子空间聚类算法都是通过各种方法来消除原始数据空间中缺失或者损坏的数据，基于一个“干净”的数据集来建立相邻关系，从而获得鲁棒的聚类效果。本文提出的基于投影的鲁棒子空间聚类算法通过在线性子空间中投影来消除误差影响，从而获得鲁棒的聚类效果，不需要噪声的先验知识。大量实验表明，该算法在人脸识别数据集上

的准确率与时间复杂度优于 LSR,LRR,SSC 等算法。在后续的工作中,可以采用更多的数据集来对本文算法进行验证和改进,可以从更多方面来考虑子空间簇的选择,以得到更好的聚类效果。在现今的大数据时代,将本文算法与大数据平台如 Hadoop 相结合来并行处理大规模数据,具有很大的研究价值。考虑到当下深度学习模型的自身优势和迅猛发展,如何把深度学习模型与本文算法相结合,开发聚类算法特征提取的新方式,也是后续的研究方向。

参 考 文 献

[1] LU L. Combined central and subspace clustering for computer vision applications[C]// International Conference on Machine Learning. ACM,2006:593-600.

[2] LEE M,CHO J,CHOI C H,et al. Procrustean Normal Distribution for Non-rigid Structure from Motion[C]// Computer Vision and Pattern Recognition. IEEE,2013:1280-1287.

[3] BOULT T E,BROWN L G. Factorization-based segmentation of motions[C]// Proceedings of the IEEE Workshop on Visual Motion. IEEE,2002:179-186.

[4] LIU D,JIANG M H,YANG X F,et al. Analyzing documents with Quantum Clustering:A novel pattern recognition algorithm based on quantum mechanics[J]. Pattern Recognition Letters,2016,77:8-13.

[5] LU S,CHANG D. An image segmentation method based on dynamic artificial fish swarm algorithm[C]// 2012 IEEE 11th International Conference on Signal Processing. IEEE,2012:980-983.

[6] ELHAMIFAR E,VIDAL R. Clustering disjoint subspaces via sparse representation[C]// IEEE International Conference on Acoustics Speech and Signal Processing. IEEE,2010:1926-1929.

[7] WU X,CHEN X M,LI X,et al. Adaptive subspace learning:an iterative approach for document clustering[J]. Neural Computing and Applications,2014,25(2):333-342.

[8] BAI J C,LI J C,DAI P F,et al. General parameterized proximal point algorithm with applications in statistical learning[J]. International Journal of Computer Mathematics,2019,96(1):199-215.

[9] WU Y,ZHANG Z,HUANG T S,et al. Multibody Grouping via Orthogonal Subspace Decomposition[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR 2001). IEEE,2001:II-252-II-257 vol. 2.

[10] YAN J,POLLEFEYS M. A General Framework for Motion Segmentation:Independent,Articulated,Rigid,Non-rigid,Degenerate and Non-degenerate[C]// European Conference on Computer Vision. Springer Berlin Heidelberg,2006:94-106.

[11] LU C Y,MIN H,ZHAO Z Q,et al. Robust and Efficient Subspace Segmentation via Least Squares Regression[C]// European Conference on Compater Vision. Berlin:Springer,2014:347-360.

[12] LIU G,LIN Z,YAN S,et al. Robust recovery of subspace structures by low-rank representation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2012,35(1):171-184.

[13] ELHAMIFAR E,VIDAL R. Sparse Subspace Clustering:Algorithm,Theory,and Applications[J]. IEEE transactions on pattern analysis and machine intelligence,2013,35(11):2765-2781.

[14] CHEN J,ZHANG H,MAO H,et al. Symmetric low-rank representation for subspace clustering[J]. Neurocomputing,2014,173(P3):1192-1202.

[15] XIA C Q,HAN K,QI Y,et al. A Self-Training Subspace Clustering Algorithm under Low-Rank Representation for Cancer Classification on Gene Expression Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2018,15(4):1315-1324.

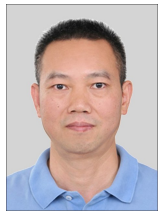
[16] LUXBURG,ULRIKE. A tutorial on spectral clustering[J]. Statistics & Computing,2007,17(4):395-416.

[17] PAPADAKIS N,BUGEAU A. Tracking with occlusions via graph cuts[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2010,33(1):144-157.

[18] HE X,YAN S,HU Y,et al. Face recognition using laplacian faces[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005:328-340.

[19] WANG J,YANG J,YU K,et al. Locality-constrained linear coding for image classification[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE,2010:3360-3367.

[20] FAVARO P,VIDAL R,RAVICHANDRAN A. A closed form solution to robust subspace estimation and clustering[C]// Computer Vision and Pattern Recognition. IEEE,2011:1801-1807.



XING Yu-hua, born in 1966, master, associate professor. His main research interests include IoT communication technology and so on.

LI Ming-xing, born in 1993, master. Her main research interests include communication of internet of things and so on.