DOI:10.11992/tis.201601005
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160718.1521.006.html

稀疏样本自表达子空间聚类算法

林大华[1],杨利锋[2],邓振云[2],李永钢[2],罗奘[2]
(1.广西电化教育馆,广西 南宁 530022; 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:针对现有子空间聚类算法在构造相似度矩阵时,没有同时利用样本自表达和稀疏相似度矩阵以及去除噪音、离群点的干扰相结合,提出了一种新的稀疏样本自表达子空间聚类方法。该方法通过样本自表达而充分利用样本间固有相关性的本质,创新性地同时使用 L_1-范数和 $L_{2,1}$-范数正则化项惩罚相似度矩阵,即对所有测试样本进行稀疏样本自表达,从而确保每个测试样本由与其相关性强的样本表示,并使所获得的相似度矩阵具有良好的子空间结构和鲁棒性。通过 Hopkins155 和人脸图像等大量数据集的实验结果表明,本文方法在实际数据的子空间聚类中能够获得非常好的效果。

关键词:子空间聚类; 谱聚类; 子空间结构; 相似度矩阵; 样本自表达
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1673-4785(2016)05-0696-07

中文引用格式:林大华,杨利锋,邓振云,等.稀疏样本自表达子空间聚类算法[J]. 智能系统学报, 2016, 11(5):696-702.
英文引用格式:LIN Dahua, YANG Lifeng, DENG Zhenyun, et al. Sparse sample self-representation for subspace clustering[J]. CAAI transactions on intelligent systems, 2016,11(5):696-702.

Sparse sample self-representation for subspace clustering

LIN Dahua[1], YANG Lifeng[2], DENG Zhenyun[2], LI Yonggang[2], LUO Yan[2]
(1.Guangxi Center for Educational Technology, Nanning 530022, China; 2. Guangxi Key Lab of Multi-source Information Mining & Security, Guilin 541004, China)

Abstract: Existing subspace clustering methods do not combine sample self-representation well with affinity matrix sparsity, for example, by removing disturbances from noise, outliers, etc., when constructing the affinity matrix. This paper proposes a novel subspace clustering method called sparse sample self-representation for subspace clustering. This method fully considers the correlation between the samples, and also takes advantage of L_1-norm and $L_{2,1}$-norm terms to "penalize" the affinity matrix; that is, it conducts sparse sample self-representation for all test samples, to guarantee every sample can be expressed by any other samples with strong similarity and make it more robust. The experimental results of the Hopkins155 dataset and some facial image datasets show that the proposed method outperforms the LSR, SSC, and LRR methods in terms of the subspace clustering error.

Keywords: subspace clustering; spectral clustering; subspace structure; similarity matrix; sample self-representation

近几年来,子空间聚类[1]方法作为一种实现高维数据聚类的有效途径,在机器学习、图像处理和计算机视觉等领域已得到广泛应用。其中基于稀疏表示(sparse representation)和低秩表示(low-rank representation)的子空间聚类方法,通过与图划分的谱聚类方法相结合,在运动图像分割、人脸识别等高维数据的聚类方面得到了较好的效果。

子空间聚类又称子空间分割是指把数据的原始特征空间分割为不同的特征子集,从不同的子空间角度考察各个样本聚类划分的意义,同时在聚类过程中为每个样本寻找相应的特征子空间。目前实现子空间聚类的方法主要归为以下 4 类:基于代数的,

收稿日期:2016-01-04.　网络出版日期:2016-07-18.
基金项目:国家自然科学基金项目(61263035, 61573270, 61450001);国家 973 计划项目(2013CB329404);中国博士后科学基金项目(2015M570837);广西自然科学基金项目(2015GXNSFCB139011);广西研究生教育创新计划项目(YCSZ2016046, YCSZ2016045).
通信作者:杨利锋. E-mail:517567113@qq.com.

如 GPCA 算法[2];基于迭代的,如 K-subspaces[3];基于统计的,如 PCA 和 RANSAC 算法[4];以及基于谱聚类的,如 SSC(sparse subspace clustering)[5],LRR(low-rank representation)[6] 和 LSR(least squares regression)[7] 算法等。其中,基于谱聚类的子空间聚类算法利用样本的局部或全局信息去构建一个相似度矩阵,然后通过谱聚类算法对样本进行聚类。该类方法能较好地处理具有噪音和离群点的数据,不需要事先知道子空间的个数以及维数,因此在手写体识别、人脸识别以及运动分割等多个应用领域获得非常好的效果。

目前比较流行的算法有基于谱聚类的 SSC 和 LRR 算法,它们主要是对每个样本都找到一组稀疏或低秩的线性表示去构建相似度矩阵,然后利用谱聚类得到最终的结果。其中,SSC 算法能够很好地结合样本自表达和稀疏相似度矩阵,通过稀疏相似度矩阵可使每个样本由与其相似性很强的一些样本表示,这些具有强相似性的样本往往在同一个子空间里,所以具有一定稀疏性的相似度矩阵往往可以提高子空间聚类的效果。但是,在数据的信噪比小、子空间不相互独立的情况时,该方法的聚类效果就不是很好。而 LRR 算法可以很好地使样本自表达和去除噪音、离群点相结合,但是其通过低秩表示构造的相似度矩阵往往不稀疏,没有很好地利用样本间的强相关性,这会影响子空间聚类的效果。

因此,合理地结合利用样本自表达和稀疏相似度矩阵以及去除噪音、离群点的干扰,能够实现构造一个良好的相似度矩阵而获得更好的子空间聚类效果的目的。所以,本文提出的方法首先从样本之间的相关性出发,对所有测试样本进行样本自表达,并同时通过 L_1-norm 和 $L_{2,1}$-norm 正则化项惩罚相似度矩阵,进行稀疏约束得到全局最优的相似度矩阵。然后,利用谱聚类得到最终的子空间聚类结果。在样本自表达过程中,L_1-norm 正则化项用来实现相似度矩阵的稀疏,确保每个测试样本都由与之相关性强的样本表示,能很好地解决样本自表达和稀疏相似度矩阵相结合的问题;而 $L_{2,1}$-norm 通过控制相似度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,最终可以实现样本自表达和稀疏相似度矩阵以及去除噪音、离群点的干扰相结合,提高子空间聚类的效果。本文将所提出的方法称为稀疏样本自表达子空间聚类算法,简称为 SSR_SC(sparse sample self-representation for subspace clustering)。

1　相关理论

高维数据一般可由多个低维结构表示,且具有很强相似度的样本往往在同一低维结构里,否则在不同的结构。每个低维结构对应为一个子空间,所以对数据的聚类可以通过对子空间的划分来聚类,即子空间聚类。

子空间聚类(subspace clustering)的定义:给定一个足够大的数据集 $\boldsymbol{X}=[x_1\cdots x_i\cdots x_n]\in\mathbf{R}^{d\times n}$,其中 x_i 表示一个具有 d 维属性的样本点,n 为样本数。假设这些样本点是分别从 k 个不同的子空间 $\{S_i\}_{i=1}^{k}$ $(i=1,2,\cdots,k)$ 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:1)根据子空间策略构造样本集的相似度矩阵 $\boldsymbol{S}$;2)通过计算相似度矩阵或拉普拉斯矩阵的前 k 个特征值与特征向量,构建特征向量空间;3)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。由上述的过程可知,该类方法的主要挑战就是构造一个良好的相似度矩阵 $\boldsymbol{S}$。而通过一个良好矩阵 $\boldsymbol{S}$ 得到的子空间的特征表现为:子空间内的样本具有高度的相似性,不同子空间的样本不相似或差异性大,且所有子空间呈块对角化结构[8]。

为了能构造一个良好的相似度矩阵而获得很好的子空间聚类效果,E. Elhamifar 等[5] 提出了稀疏子空间聚类算法 SSC(sparse subspace clustering),将每个样本用空间里的其他样本来线性表示,通过 L_1-norm 正则化项确保所获得的相似度矩阵是稀疏的。而 Liu 等[6] 提出低秩自表达 LRR(low-rank representation),通过核范数利用数据的全局信息寻找低秩样本自表达,并用 $L_{2,1}$-norm 项去除噪音和离群点,使所构造的相似度矩阵具有鲁棒性。由 Lu 等[7] 提出的 LSR(least squares regression),通过优化求解最小二乘回归的目标函数 $\min\limits_{Z}\|\boldsymbol{X}-\boldsymbol{XZ}\|_F^2+\lambda\|\boldsymbol{Z}\|_F^2$,能获得具有良好块对角化结构的相似度矩阵。但是,这些方法都没有同时利用样本自表达和稀疏相似度矩阵以及去除噪音、离群点的干扰相结合来构造相似度矩阵。

因此,本文提出了 SSR_SC 算法,充分利用样本间的相关性进行样本自表达,并通过 L_1-norm 和 $L_{2,1}$-norm 正则化项对相似度矩阵进行稀疏约束,从而能很好地实现构造一个良好的相似度矩阵的目的。

2　SSR_SC 算法

对于样本空间 $\boldsymbol{X}$ 中的一个样本 x,用同一空间中的其他样本对 x 进行线性表示的过程称为样本自表达。同一子空间中的样本之间往往具有很强的相关性,不同子空间的样本之间为无相关性或弱相关

性,所以通过样本自表达能很好地利用样本之间的相关性来提高子空间聚类的效果。假设样本空间为 $\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\ \boldsymbol{x}_n]\in\mathbf{R}^{d\times n}$,其中 n 为样本数,$\boldsymbol{x}_i(i=1,2,\cdots,n)$ 为具有 d 维属性的样本点。根据上述样本自表达的定义,即找出这样一个列向量 $\boldsymbol{z}_i\in\mathbf{R}^{n\times 1}$,使得 $\boldsymbol{x}_i$ 可以通过 $\boldsymbol{Xz}_i$ 表示。

但是样本空间 $\boldsymbol{X}$ 中往往存在噪音和离群点等干扰,其用 $\boldsymbol{e}$ 表示,则 $\boldsymbol{x}_i=\boldsymbol{Xz}_i+\boldsymbol{e}$,而这些干扰通常会影响子空间聚类的效果。因此本文算法希望找到这样一个相似度矩阵 $\boldsymbol{Z}=[\boldsymbol{z}_1\ \boldsymbol{z}_2\cdots\ \boldsymbol{z}_n]\in\mathbf{R}^{n\times n}$,使得 $\boldsymbol{X}$ 与 $\boldsymbol{XZ}$ 的误差尽可能小。这通常可采用岭回归(ridge regression)实现:

$$\min_{\boldsymbol{Z}}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{Xz}_i\|_2^2+\|\boldsymbol{Z}\|_2^2=\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{XZ}\|_F^2+\lambda\|\boldsymbol{Z}\|_2^2,\text{s.t. diag}(\boldsymbol{Z})=0 \tag{1}$$

式中:$\|\cdot\|_F^2$ 为 Frobenius 范数,根据岭回归求解可得 $\boldsymbol{Z}^*=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda\boldsymbol{I}_n)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$。但是这样得到的矩阵 $\boldsymbol{Z}$ 通常不稀疏,并且不能很好地解决噪音和离群点的干扰。因此本文算法利用 L_1-norm 和 $L_{2,1}$-norm 替代目标函数(1)中 L_2-norm,得到如下目标函数:

$$\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{XZ}\|_F^2+\lambda_1\|\boldsymbol{Z}\|_1+\lambda_2\|\boldsymbol{Z}\|_{2,1},\quad\text{s.t. diag}(\boldsymbol{Z})=0 \tag{2}$$

式中:$\|\boldsymbol{Z}\|_1=\sum_{i=1}^{n}\sum_{j=1}^{n}|z_{ij}|$,$\lambda_1$ 作为 L_1-norm 的参数,用于控制矩阵 $\boldsymbol{Z}$ 整体的稀疏性[9],使得每个样本 $\boldsymbol{x}_i$ 都由与之具有强相关性的样本表示;$\|\boldsymbol{Z}\|_{2,1}=\sum_{i=1}^{n}(\sum_{j=1}^{n}|z_{ij}|^2)^{\frac{1}{2}}$,而 λ_2 作为 $L_{2,1}$-norm 的参数,用于调整矩阵 $\boldsymbol{Z}$ 整行的稀疏性,可以去除噪音和离群点的干扰[10]。此时,通过该目标函数构造获得的相似度矩阵,能很好地同时利用样本自表达和稀疏相似度矩阵以及去除噪音、离群点的干扰相结合,使得子空间聚类的效果更好。

假定通过目标函数(2)得到如下矩阵 $\boldsymbol{Z}$,其中 $\text{diag}(\boldsymbol{Z})=0$,表明样本不能将自身作为相关性样本进行线性表示。

$$\boldsymbol{Z}=\begin{bmatrix}0 & 0.8 & 0.1 & 0.7\\ 0.6 & 0 & 0 & 0.4\\ 0 & 0 & 0 & 0\\ 0.5 & 0.9 & 0 & 0\end{bmatrix}$$

根据样本自表达定义可知,矩阵 $\boldsymbol{Z}$ 的第一列即样本 $\boldsymbol{x}_1$,可以通过样本 $\boldsymbol{x}_2$ 和样本 $\boldsymbol{x}_4$ 线性表示,即 $\boldsymbol{x}_1=(\boldsymbol{x}_2\ \boldsymbol{x}_4)\times\begin{pmatrix}z_{21}\\ z_{41}\end{pmatrix}$,同理 $\boldsymbol{x}_2=(\boldsymbol{x}_1\ \boldsymbol{x}_4)\times\begin{pmatrix}z_{12}\\ z_{42}\end{pmatrix}$,$\boldsymbol{x}_3=(\boldsymbol{x}_1)\times(z_{13})$,$\boldsymbol{x}_4=(\boldsymbol{x}_1\ \boldsymbol{x}_2)\times\begin{pmatrix}z_{14}\\ z_{24}\end{pmatrix}$。其中矩阵 $\boldsymbol{Z}$ 的第 3 行全为 0,这是 $L_{2,1}$-norm 对矩阵 $\boldsymbol{Z}$ 进行行稀疏的效果,表明对样本 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_4$ 而言,$\boldsymbol{x}_3$ 为不相关样本即噪音样本。

根据以上分析可知,本文提出的目标函数(2),不仅通过 L_1-norm 确保每个样本都由与之具有强相关性的样本表示,而且利用 $L_{2,1}$-norm 避免噪音和离群点的干扰,使其具有更好的鲁棒性。注意,此时得到的相似度矩阵 $\boldsymbol{Z}$ 是稀疏的且充分考虑了样本间的相关性以及去除了噪音和离群点的干扰,然后通过矩阵 $\boldsymbol{Z}$ 构造矩阵 $\boldsymbol{S}=(|\boldsymbol{Z}|+|\boldsymbol{Z}^{\mathrm{T}}|)/2$,接着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了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的特征,很好地解决了基于谱聚类的子空间聚类问题中构造一个良好相似性矩阵的问题。

本文提出的 SSR_SC 算法的具体步骤如下:

算法 1 SSR_SC 算法

输入 参数 λ_1 和 λ_2,以及样本空间:$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\ \boldsymbol{x}_n]\in\mathbf{R}^{d\times n}$。

输出 聚类结果 $\boldsymbol{C}\in\mathbf{R}^{n\times 1}$。

1)根据算法 2,迭代求解目标函数 $\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{XZ}\|_F^2+\lambda_1\|\boldsymbol{Z}\|_1+\lambda_2\|\boldsymbol{Z}\|_{2,1}$,获得其最优化矩阵 $\boldsymbol{Z}$;

2)根据样本相似度矩阵 $\boldsymbol{Z}$ 去构造矩阵 $\boldsymbol{S}=(|\boldsymbol{Z}|+|\boldsymbol{Z}^{\mathrm{T}}|)/2$;

3)利用谱聚类算法得到最终的聚类结果 $\boldsymbol{C}\in\mathbf{R}^{n\times 1}$。

3 目标函数优化

目标函数(2)是一个凸函数,但是 L_1-norm 和 $L_{2,1}$-norm 是非光滑的,无法直接求得解析解。为此,本文提出了一种有效的优化算法来解决这个问题,最后解出目标函数的最优化结果。

对目标函数(2)中的 $\boldsymbol{Z}$ 的每一行 $\boldsymbol{Z}_i$ 求导,然后令其为 0,得到式(3)

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XZ}_i-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda_1\boldsymbol{D}_i\boldsymbol{Z}_i+\lambda_2\tilde{\boldsymbol{D}}\boldsymbol{Z}_i=0 \tag{3}$$

式中:$\boldsymbol{D}_i(1\leqslant i\leqslant n)$ 是一个对角矩阵,其第 k 个对角元素为 $\frac{1}{2|\boldsymbol{Z}_{ki}|}$;$\tilde{\boldsymbol{D}}$ 也是一个对角矩阵,其第 k 个对角元素为 $\frac{1}{2\|\boldsymbol{Z}^k\|_2}$。因此可以将式(3)表示为

$$\boldsymbol{Z}_i=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda_1\boldsymbol{D}_i+\lambda_2\tilde{\boldsymbol{D}})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \tag{4}$$

此时,数据集 $\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{X}$ 已知,λ_1 和 λ_2 为调制参

数。但值得注意的是,$\boldsymbol{D}_i$ 和$\tilde{\boldsymbol{D}}$依赖于 $\boldsymbol{Z}$,因此也是未知的。根据文献[11-13],本文提出了一种迭代算法去求解这个问题。

算法 2　目标函数优化算法

输入　数据集 $\boldsymbol{X}$;

输出　$\boldsymbol{Z}^{(t)} \in \mathbf{R}^{n\times n}$;

初始化$\boldsymbol{Z}^{(1)} \in \mathbf{R}^{n\times n}$,$t=1$;

do{

1)计算对角矩阵$\boldsymbol{D}_i^{(t)}$($1 \leqslant i \leqslant n$)和$\tilde{\boldsymbol{D}}^{(t)}$其中$\boldsymbol{D}_i^{(t)}$的第 k 个对角元素为$\dfrac{1}{2|\boldsymbol{Z}_{\boldsymbol{ki}}^{(t)}|}$,$\tilde{\boldsymbol{D}}^{(t)}$的第 k 个对角元素为$\dfrac{1}{2\|(\boldsymbol{Z}^{(t)})^k\|_2}$;

2)For 每个 $i(1\leqslant i\leqslant n)$,计算

$$\boldsymbol{Z}_i^{(t+1)} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda_1 \boldsymbol{D}_i^{(t)} + \lambda_2 \tilde{\boldsymbol{D}}^{(t)})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$$

3)$t=t+1$;

}until 收敛。

由上述算法 2 的描述知,优化目标函数的关键是目标值$\boldsymbol{Z}_i^{(t+1)}$ 要在迭代中收敛。下面证明目标函数(2)的目标值在每一次迭代中都会减小。根据算法 2 的第 2 步,可得

$$\boldsymbol{Z}^{(t+1)} = \min_{\boldsymbol{Z}} \mathrm{Tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}) + \lambda_1 \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{D}_i^{(t)} \boldsymbol{Z}_i + \lambda_2 \mathrm{Tr}\,\boldsymbol{Z}^{\mathrm{T}} \tilde{\boldsymbol{D}}^{(t)} \boldsymbol{Z} \tag{5}$$

由式(5),可得

$$\mathrm{Tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t+1)}) + \lambda_1 \sum_{i=1}^{n} (\boldsymbol{Z}_i^{(t+1)})^{\mathrm{T}} \boldsymbol{D}_i^{(t)} \boldsymbol{Z}_i^{(t+1)} + \lambda_2 \mathrm{Tr}(\boldsymbol{Z}^{(t+1)})^{\mathrm{T}} \tilde{\boldsymbol{D}}^{(t)} \boldsymbol{Z}^{(t+1)} \leqslant \mathrm{Tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t)})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t)}) + \lambda_1 \sum_{i=1}^{n} (\boldsymbol{Z}_i^{(t)})^{\mathrm{T}} \boldsymbol{D}_i^{(t)} \boldsymbol{Z}_i^{(t)} + \lambda_2 \mathrm{Tr}(\boldsymbol{Z}^{(t)})^{\mathrm{T}} \tilde{\boldsymbol{D}}^{(t)} \boldsymbol{Z}^{(t)}$$

于是,可推导出

$$\mathrm{Tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t+1)}) + \lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{n} \|\boldsymbol{Z}_{ij}^{(t+1)}\| + \lambda_2 \sum_{k=1}^{d} \|(\boldsymbol{Z}^{(t+1)})^k\|_2 \leqslant \mathrm{Tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t)})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}^{(t)}) \lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{n} (\|\boldsymbol{Z}_{ij}^{(t)}\|) + \lambda_2 \sum_{k=1}^{d} \|(\boldsymbol{Z}^{(t)})^k\|_2$$

根据文献[14],对于任意向量 $\boldsymbol{Z}$ 和 $\boldsymbol{Z}_0$,有 $\|\boldsymbol{Z}_2\| - \dfrac{\|\boldsymbol{Z}\|_2^2}{2\|\boldsymbol{Z}_0\|_2} \leqslant \|\boldsymbol{Z}_0\|_2 - \dfrac{\|\boldsymbol{Z}_0\|_2^2}{2\|\boldsymbol{Z}_0\|_2}$。由以上分析可知,算法 2 中的目标值在每一次迭代中都会减小。$\boldsymbol{Z}^{(t)}$、$\boldsymbol{D}_i^{(t)}$($1\leqslant i\leqslant n$)和$\tilde{\boldsymbol{D}}^{(t)}$在收敛处满足等式(5),且目标函数(2)是凸函数,于是此时得到的 $\boldsymbol{Z}$ 是式(3)的一个全局最优解,因此,算法 2 可将问题(3)收敛到全局最优解。另外,在每一次迭代时都有封闭形式的解,因此我们的算法收敛非常快。

4　实验分析与讨论

4.1　实验数据集及评价指标

本文算法通过 MATLAB 语言编程,且所有实验都是在 win7 系统下的 MATLAB 2014 软件上运行测试。实验用到的数据集介绍如下。

Hopkins155[15]数据集被广泛用来测试各种子空间聚类算法。该数据集由 156 个视频序列组成,一个序列对应一个数据集,所以其共有 156 个数据集,并且每个序列中包含 2 或 3 个运动物体目标。

Jaffe[16]数据集由日本 ATR 表情识别研究协会提供,该数据集包含 10 个人的 213 张表情图像,每张表情图像经过预处理被裁剪为 32 像素×32 像素大小的尺度。

USPS[17]数据集是由美国国家邮政局提供,数据集含有 9 298 个 0~10 的手写数字数据集,每个手写数字数据经预处理都被裁剪为 16 像素×16 像素大小的尺度。用每个数字的前 100 个图像进行实验。

ORL[18]数据集是由剑桥 Olivetti 实验室提供,数据集包含 40 人的共 400 张面部图像,每张人脸数据经预处理被裁剪为 16 像素×16 像素大小的尺度。

为了验证算法的性能,将目前较好的子空间聚类算法 LSR、LRR 和 SSC 与本文算法进行对比实验。为了保证算法的公平性,所有算法都没有对数据进行后期处理。

子空间聚类的重要挑战是处理存在于数据中的错误。因此,本文将子空间聚类错误率作为衡量各个算法性能的评价标准。其中,错误率越小,子空间聚类效果越好;反之,则越差。其定义为

$$\text{子空间聚类错误率} = \frac{\text{错误分类的样本数}}{\text{样本总数}}$$

4.2　实验结果与分析

4.2.1　Hopkins155 数据集上的实验

由于 Hopkins155 数据集包含 156 个不同的数据集,根据文献[19],本文将 156 个数据集中的子空间聚类错误率的最大值(Max)、均值(Mean)和中值(Median)以及标准差(Std)作为评价指标。对 LSR,LRR,SSC 以及本文算法 SSR_SC 在该数据集进行了对比,实验结果如表 1 所示。

通过分析表 1 可知,在 Hopkins155 数据集上,本文提出的 SSR_SC 比 LSR、LRR 和 SSC 获得了更好的子空间聚类效果。具体地,SSR_SC 与 LSR 算

法对比,错误率均值小 2.38%,标准差小 3.12%。LSR 中使用 L_2-norm 正则化项约束相似度矩阵 $\boldsymbol{Z}$,能使 $\boldsymbol{Z}$ 具有很好的块对角化结构,但是其并没有对 $\boldsymbol{Z}$ 稀疏而影响其最终的聚类效果。SSR_SC 与 LRR 算法对比,最大错误率小 7.16%,均值小 3.30%,标准差小 4.53%。其中 LRR 利用 $L_{2,1}$-norm 项惩罚相似度矩阵 $\boldsymbol{Z}$ 而可以去除噪音和离群点的影响,但是,其没有对 $\boldsymbol{Z}$ 稀疏。而本文提出的 SSR_SC 算法通过 $L_{2,1}$-norm 正则化项惩罚相似度矩阵而使其具有鲁棒性,且还对 $\boldsymbol{Z}$ 进行稀疏,所以能获得更好的子空间聚类效果。与 SSC 算法比较,本文算法 SSR_SC 也取得了更好的效果,最大错误率小 0.96%,均值错误率小 1.19%,标准差错误率小 2.14%。Hopkins155 数据集的大部分数据都是比较干净的,只有很小部分数据受到污染,这样的条件下 SSC 稀疏$\boldsymbol{Z}$而更充分地利用了样本间的强相关性,从表 1 中可以看到 SSC 比 LRR 的子空间聚类效果更好。

表 1 LSR、LRR、SSC 和 SSR_SC 在 Hopkins155 数据集上实验的子空间聚类错误率

Table 1 Subspace clustering errors of LSR, LRR, SSC and SSR_SC on Hopkins155 dataset

错误率	LSR	LRR	SSC	SSR_SC
最大值	0.397 1	0.476 4	0.414 4	0.404 8
平均值	0.042 1	0.051 3	0.030 2	0.018 3
中位数	0.005 2	0.005 3	0	0
标准差	0.086 0	0.100 1	0.076 2	0.054 8

4.2.2 数字图像和人脸图像上的实验

为了证明本文算法 SSR_SC 在实际数据集中也具有适用性,本文还在 USPS、ORL 以及 Jaffe 等数字图像和人脸图像数据集也进行了对比实验。实验结果如表 2 所示。

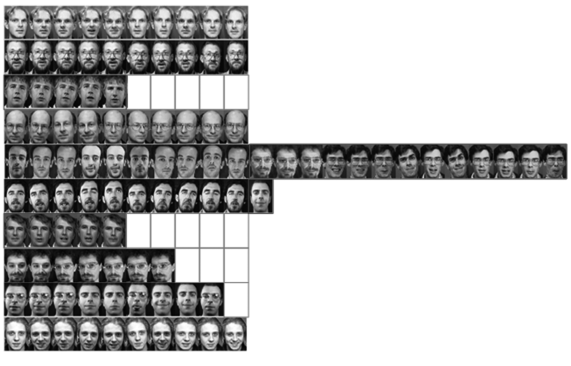
表 2 LSR、LRR、SSC 和 SSR_SC 分别在 Jaffe 和 ORL 以及 USPS 数据集实验的子空间聚类错误率

Table 2 Subspace clustering errors of LSR, LRR, SSC and SSR_SC on Jaffe, ORL and USPS dataset

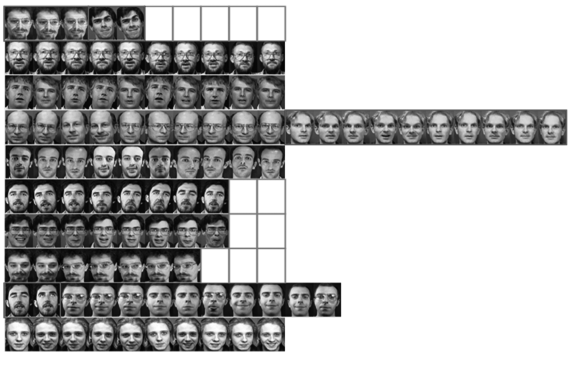
数据集	LSR	LRR	SSC	SSR_SC
USPS	0.261 0	0.367 0	0.475 0	0.124 0
ORL	0.222 5	0.550 0	0.225 0	0.207 5
Jaffe	0.379 1	0.136 2	0.117 4	0.014 1

从表 2 数据可知,SSR_SC 算法在 Jaffe 数据集上取得了最优的效果,其子空间聚类错误率为 1.41%,远远低于 LSR,LRR 和 SSC 算法的错误率,效果与 LRR 和 SSC 相比提升了 10 倍,甚至比 LSR 算法提高了接近 40 倍。而在 USPS 和 ORL 数据集上同样也取得了较低的子空间聚类错误率,其中在 USPS 数据集上,相比 LSR、LRR 和 SSC 分别提高了 13.70%、24.30%、35.10%;在 ORL 数据集上分别提高了 1.50%、34.25%、1.75%。因此,可以认为本文提出的 SSR_SC 算法是一种高效的子空间聚类算法。

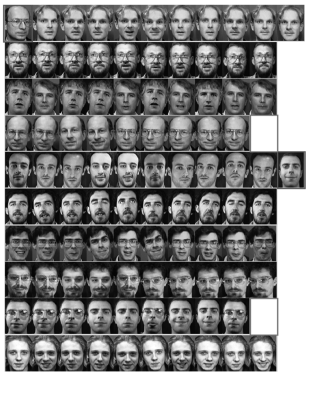
为了更加直观地对比 LRR、SSC 和 SSR_SC 算法的子空间聚类效果,选取 ORL 数据集里 100 张图片(10 人,每人 10 张)进行实验,得到的实验结果如图 1 所示。



(a)SSC



(b)LRR



(c)SSR_SC

图 1 SSC 算法、LRR 算法和 SSR_SC 算法分别对 ORL 数据集聚类的效果

Fig.1 The subspace clustering effect of SSC, LRR and SSR_SC on ORL dataset

图 1 中,每行都代表一个子空间,短划线方框区域表示错误聚类的图片。从图 1 可以直观的看出,本文提出的 SSR_SC 算法取得的子空间聚类效果明显好于 LRR 算法和 SSC 算法。其中,SSR_SC 只将该数据集的 2 个人错误地聚类到其他子空间,而 LRR 和 SSC 算法聚类错误的图片数量分别为 17 张和 19 张,其中还存在将同一个人的图像平均的聚类为 2 个子空间的情况,如图 1(a)方点线方框所示,甚至出现将不同 2 组人聚类到同一个子空间的情况。综上分析,SSR_SC 算法比现有的子空间聚类方法有更好的子空间聚类效果。

5　结束语

提出一种综合稀疏学习和样本自表达的子空间聚类方法称为稀疏样本自表达算法。该算法通过充分考虑样本之间的相关性而进行样本自表达,并且通过稀疏学习理论进行优化,即通过 L_1-norm 使相似度矩阵得到适当稀疏而让每个样本由与其相似性高的样本进行表达,通过 $L_{2,1}$-norm 解决样本自表达过程中噪音和离群点的干扰。与 SSC 算法和 LRR 算法比较,SSR_SC 算法具有更好的鲁棒性和实现构造一个良好相似度矩阵的目的。此外,在 Hopkinss155、USPS、ORL 和 Jaffe 等数据集上实验的结果表明,SSR_SC 算法在实际数据集,如运动目标分割和图像聚类等方面,能获得更好的子空间聚类效果。此后工作将提出的方法应用于更广泛的领域,如医学数据、文本数据以及金融数据等高维数据的聚类分析。

参考文献:

[1]王卫卫, 李小平, 冯象初, 等. 稀疏子空间聚类综述[J]. 自动化学报, 2015, 41(8): 1373-1384.
WANG Weiwei, LI Xiaoping, FENG Xiangchu, et al. A survey on sparse subspace clustering[J]. Acta automatica sinica, 2015, 41(8): 1373-1384.

[2]VIDAL R, MA Yi, SASTRY S. Generalized principal component analysis (GPCA)[J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(12): 1945-1959.

[3]TSENG P. Nearest q-flat to m points[J]. Journal of optimization theory and applications, 2000, 105(1): 249-252.

[4]FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.

[5]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.

[6]LIU Guangcan, LIN Zhouchen, YAN Shuicheng, 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.

[7]LU Canyi, MIN Hai, ZHAO Zhongqiu, et al. Robust and efficient subspace segmentation via least squares regression [C]//Proceedings of the 12th European Conference on Computer Vision. Berlin Heidelberg, 2012: 347-360.

[8]FENG Jiashi, LIN Zhouchen, XU Huan, et al. Robust subspace segmentation with block-diagonal prior[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 3818-3825.

[9]欧阳佩佩, 赵志刚, 刘桂峰. 一种改进的稀疏子空间聚类算法[J]. 青岛大学学报: 自然科学版, 2014, 27(3): 44-48.
OUYANG Peipei, ZHAO Zhigang, LIU Guifeng. An improved sparse subspace clustering algorithm[J]. Journal of Qingdao university: natural science edition, 2014, 27(3): 44-48.

[10]杨国亮, 罗璐, 丰义琴, 等. 基于低秩稀疏图的结构保持投影算法[J]. 计算机工程与科学, 2015, 37(8): 1584-1590.
YANG Guoliang, LUO Lu, FENG Yiqin, et al. Structure preserving projection algorithm based on low rank and sparse graph [J]. Computer engineering and science, 2015, 37(8): 1584-1590.

[11]ZHU Xiaofeng, HUANG Zi, YANG Yang, et al. Self-taught dimensionality reduction on the high-dimensional small-sized data[J]. Pattern recognition, 2013, 46(1): 215-229.

[12]ZHANG Shichao, QIN Zhenxing, LING C X, et al. "Missing is useful": missing values in cost-sensitive decision trees[J]. IEEE transactions on knowledge and data engineering, 2005, 17(12): 1689-1693.

[13]ZHU Xiaofeng, HUANG Zi, CUI Jiangtao, et al. Video-to-shot tag propagation by graph sparse group lasso[J]. IEEE transactions on multimedia, 2013, 15(3): 633-646.

[14]ZHU Xiaofeng, ZHANG Lei, HUANG Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE transactions on image processing, 2014, 23(9): 3737-3750.

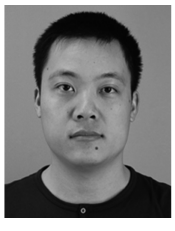
[15]TRON R, VIDAL R. A benchmark for the comparison of 3-D motion segmentation algorithms [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA, 2007: 1-8.

[16]LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with gabor wavelets[C]//Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition. Nara, 1998: 200-205.

[17]HULL J J. A database for handwritten text recognition re-

search[J]. IEEE transactions on pattern analysis and machine intelligence, 1994, 16(5): 550-554.
[18] SAMARIA F S, HARTERT A C. Parameterisation of a stochastic model for human face identification[C]//Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision. Sarasota, FL, USA, 1994: 138-142.
[19] FENG Jiashi, LIN Zhouchen, XU Huan, et al. Robust subspace segmentation with block-diagonal prior[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 3818-3825.

作者简介：



林大华，男，1979 年生，主要研究方向为机器学习、数据挖掘。

杨利锋，男，1989 年生，硕士研究生，主要研究方向为数据挖掘和机器学习。

邓振云，男，1991 年生，硕士研究生，主要研究方向为机器学习、数据挖掘。发表学术论文 8 篇，其中被 SCI、EI 检索 4 篇。

2017 国际群体智能会议
The Eighth International Conference on Swarm Intelligence (ICSI'2017)

July 27-August 01, 2017, Fukuoka, Japan

The Eighth International Conference on Swarm Intelligence (ICSI'2017) serves as an important forum for researchers and practitioners to exchange latest advantages in theories, technologies, and applications of swarm intelligence and related areas. The ICSI'2017 is the eighth annual event in this high-reputation ICSI series after Bali event (ICSI'2016), Beijing joint event (ICSI-CCI'2015), Hefei event (ICSI'2014), Harbin event (ICSI'2013), Shenzhen event (ICSI'2012), Chongqing event (ICSI'2011) and Beijing event (ICSI'2010). Papers presented at the ICSI'2017 will be published in Springer's Lecture Notes in Computer Science (indexed by EI Compendex, ISTP, DBLP, SCOPUS, Web of Science ISI Thomson, etc.), some high-quality papers will be selected for SCI-indexed Transaction and Journal (including IEEE/ACM Transactions on CBB, NC, CC, IJSIR, IJCIPR, etc.).

The ICSI'2017 will be held in the center of the Fukuoka City. Historical city, Fukuoka, is the 5th largest city in Japan with 1.55 million populations and is the 7th most liveable city in the world according to the 2016 Quality of Life Survey by Monocle. Fukuoka locates at the northern end of the Kyushu Island and is the economic and cultural center of whole Kyushu Island. Because of its closeness to the Asian mainland, Fukuoka has been an important harbor city for many centuries. Today's Fukuoka is the product of the fusion of two cities in the year 1889, when the port city of Hakata and the former castle town of Fukuoka were united into one city called Fukuoka.

Website: http://www.ic-si.org/