局部子空间聚类

刘展杰 [1] 陈晓云 [1]

摘 要 现有子空间聚类方法通常以数据全局线性为前提, 将每个样本点表示为其他样本点的线性组合, 因而导致常见子空间聚类方法不能很好地应用于非线性数据. 为克服全局线性表示的局限, 借鉴流形学习思想, 用 k 近邻局部线性表示代替全局线性表示, 与稀疏子空间聚类和最小二乘子空间聚类方法相结合, 提出局部稀疏子空间聚类和局部最小二乘子空间聚类方法, 统称局部子空间聚类方法. 在双月形数据、6 个图像数据集和 4 个基因表达数据集上进行实验, 实验结果表明该方法是有效的.

关键词 局部线性, k 近邻, 子空间聚类, 图像数据, 基因表达数据

引用格式 刘展杰, 陈晓云. 局部子空间聚类. 自动化学报, 2016, **42**(8): 1238−1247

DOI 10.16383/j.aas.2016.c150335

Local Subspace Clustering

LIU Zhan-Jie[1] CHEN Xiao-Yun[1]

Abstract Existing subspace clustering methods usually rest on a global linear data set, which expresses each data point as a linear combination of all other data points, and thus common methods are not well suited for the nonlinear data. To overcome this limitation, the local sparse subspace clustering and local least squares regression subspace clustering are proposed. The idea of the two new methods comes from manifold learning which expresses each data point as a linear combination of its k nearest neighbors, and is combined with sparse subspace clustering and least squares subspace clustering respectively. Experimental results show that our method is effective on two-moon synthetic data, six image data sets and four gene expression data sets.

Key words Local linear, k nearest neighbors, subspace clustering, image data, gene expression data

Citation Liu Zhan-Jie, Chen Xiao-Yun. Local subspace clustering. *Acta Automatica Sinica*, 2016, **42**(8): 1238−1247

许多现实采样的数据特征复杂, 呈非线性分布, 变量之间存在高度非线性相关性. 大部分情况下, 图像数据被视为一种具有非线性结构的数据. 因图像易受外界因素的干扰, 使得数据具有很强的不确定性. 例如人脸图像会受光照、表情和姿态等因素的影响, 通常导致图像中的像素值发生非线性的变化. 同样, 现实中采集到的基因表达数据也常呈现出高维、小样本、多噪声和非线性等特征, 且大部分数据不含类别信息, 因此聚类技术作为一种典型无监督技术而被广泛使用. 故本文从非线性角度研究图像数据和基因表达数据的聚类问题.

近年来, 子空间聚类作为一种热点聚类方法被广泛研究, 并被应用到机器视觉领域, 如图像分割领域[1]、动态分割领域[2] 和人脸聚类领域[3]; 此外图像表达[4] 和混合识别系统[5] 等图像处理领域中, 子空间聚类方法也被成功使用. 目前, 已有许多经典的子空间聚类算法被提出, 如稀疏子空间聚类 (Sparse subspace clustering, SSC)[6]、低秩表达 (Low rank representation, LRR)[7] 和最小二乘子空间聚类 (Least squares regression, LSR)[8] 等, 三者都是基于谱聚类的子空间聚类算法. SSC 可实现从全局角度自动选择近邻点, 消除数据之间的相关性; LRR 保持仿射矩阵的低秩性; LSR 保持数据的聚集性. 在这些方法基础之上改进得到鲁棒潜在低秩表达子空间聚类 (Robust latent low rank representation, RLLRR)[9]、光滑表达子空间聚类 (Smooth representation clustering, SMR)[10]、鲁棒 SSC (Robust SSC, R-SSC)[11]、块对角 SSC (Block-diagonal SSC, BD-SSC) 和块对角 LRR (Block-diagonal LRR, BD-LRR)[12] 等.

上述的大部分方法都以数据全局线性为前提, 每个样本点可以利用其余样本点线性重构. 从回归模拟角度来看, 上述方法可看作线性回归模型. 故 SSC 是 LASSO (Least absolute shrinkage and selectionator operator)[13] 的变形, LSR 是岭回归

收稿日期 2015-05-29 录用日期 2015-11-26
Manuscript received May 29, 2015; accepted November 26, 2015
国家自然科学基金 (71273053, 11571074), 福建省自然科学基金 (2014J01009) 资助
Supported by National Natural Science Foundation of China (71273053, 11571074) and Natural Science Foundation of Fujian Province (2014J01009)
本文责任编委 周志华
Recommended by Associate Editor ZHOU Zhi-Hua
1. 福州大学数学与计算机科学学院 福州 350116
1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116

(Ridge regression)[14] 的变形. LASSO 以 l_1 范数作为惩罚项来压缩模型回归系数, l_1 范数在原点处不可导, 通常采用迭代方法求解模型的回归系数. 因求解会得到许多为 0 的回归系数, 所对应的回归变量在线性回归模型中不起作用, 以此达到压缩变量的目的[15]. 但对有较多变量的模型来说, 压缩变量的效果会变差. 岭回归是一种较稳定的连续型模型, 但它不能压缩变量, 这导致模型中变量过多, 不能得到一个简单且解释性强的模型. 对于线性相关程度大的数据, LASSO 和岭回归的效果较好; 反之, 这两种方法的回归效果较差. 在子空间聚类方法中, SSC 和 LSR 也保持着同样的优缺点. 因此对于非线性数据, SSC 和 LSR 的聚类效果不理想, 如图 1 所示.

图 1 是 LSR 和 SSC 在双月形数据上学习得到的仿射矩阵画出的邻接图. 由图 1 (a) 可见, LSR 用除本身外的其余点做线性表示, 不能将两类数据区分开. 由图 1 (b) 中亦可见, SSC 虽然能自动选取样本点进行线性表示, 但样本点的选取不理想. 可见 LSR 和 SSC 都不能达到图 1 (c) 的理想效果, 二者不能够将两类数据准确地区分开, 因此不适合对非线性数据进行聚类.

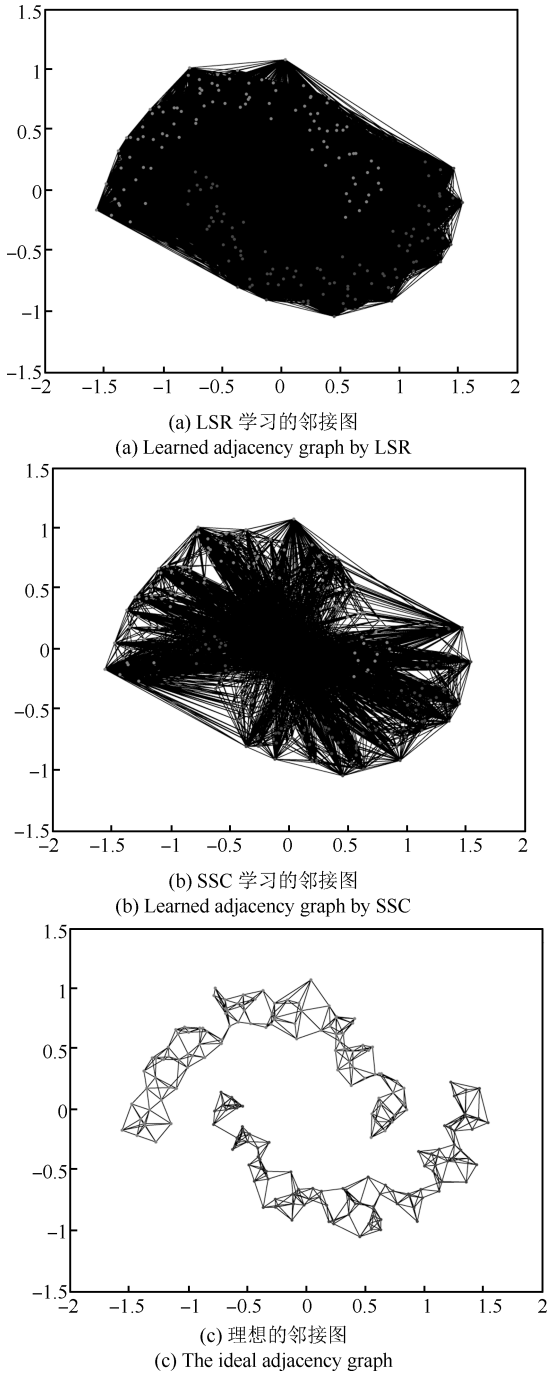
流形学习是近年来处理流形数据和非线性数据的常用方法, 许多经典流形学习方法被提出, 如局部保持投影 (Locality preserving projections, LPP)[16]、局部线性嵌入 (Locally linear embedding, LLE)[17] 和近邻保持嵌入 (Neighborhood preserving embedding, NPE)[18] 等. LPP、LLE 和 NPE 有一个共同点就是假设数据是局部线性的, 每个样本点均可利用其近邻样本进行线性表示. 同样, 在聚类问题中, 相互靠近的样本点往往被视为同类, 这样便可以利用近邻点对数据线性表示. 因此, 受流形学习局部线性表示思想的启发, 结合子空间聚类, 本文提出局部子空间聚类算法 (Local subspace clustering, LSC), 并应用于图像和基因表达谱等具有明显非线性特征的高维数据.

本文提出的局部子空间聚类利用样本点的 k 个近邻进行线性表示. 值得注意的是: 1) k 近邻法的关键是近邻参数 k 的选取, 本文提出的子空间聚类方法能够自适应地调整近邻点个数. 2) 通常采用 k 近邻法只选取适当少的近邻点, 但局部子空间聚类算法既可以取适当少, 也可以取适当多的近邻点. 当选取适当少的近邻点时, 可剔除大量样本点, 计算便捷; 当选取适当多的近邻点时, 可剔除一些噪声点, 使算法更具鲁棒性.

1 子空间聚类

子空间聚类目标是将样本数据分割或聚集成几个簇, 每个簇对应一个子空间, 因此也称为子空间分割.

定义 1[8]. 子空间聚类 (Subspace clustering): 给定的从 c 个子空间 $\{S_i\}_{i=1}^{c}$ 采样的数据向量集合 $X = [X_1, X_2, \cdots, X_c] = [x_1, x_2, \cdots, x_n] \in \mathbf{R}^{d\times n}$, X_i 是从子空间 S_i 中采样的 n_i 个数据向量构成的集合, $n = \sum_{i=1}^{c} n_i$. 子空间聚类的目标是根据它们采样的基子空间分割数据.



(a) LSR 学习的邻接图
(a) Learned adjacency graph by LSR

(b) SSC 学习的邻接图
(b) Learned adjacency graph by SSC

(c) 理想的邻接图
(c) The ideal adjacency graph

图 1 在双月形数据上学习的邻接图

Fig. 1 Learned adjacency graph on the two-moon synthetic data

子空间聚类算法根据表示方式不同可以粗略分为四类[8]: 代数方法、迭代方法、统计方法和谱聚类方法, 本文重点研究最后一种. 基于谱聚类的子空间聚类方法核心在于求解仿射矩阵 $Z=(z_{ij})_{n\times n}$, z_{ij} 用来度量两个样本 x_i 和 x_j 的相似度. 典型的相似度度量 $z_{ij}=\exp(-\|x_i-x_j\|^2/\sigma), \sigma>0$ 不能很好地刻画数据的本质特征[8]. 稀疏子空间聚类 SSC、低秩表达子空间聚类 LRR 和最小二乘回归子空间聚类 LSR 提出了新的仿射矩阵计算方法. 这些方法将每个样本点 x_i 表示为其他样本点的线性组合:

$$x_i=\sum_{j\neq i}x_j z_{ij} \tag{1}$$

其中 z_{ij} 是表示权重的常数. 利用表示系数 $(|z_{ij}|+|z_{ji}|)/2$ 度量 x_i 和 x_j 之间的相似度.

SSC 的目标函数:

$$\begin{aligned}&\min_z \|Z\|_1\\ &\text{s.t.}\quad X=XZ,\quad \text{diag}\{Z\}=0\end{aligned} \tag{2}$$

其中, $\|Z\|_1$ 是 Z 的 l_1 范数, 定义为仿射矩阵 Z 的所有元素的绝对值之和. 文献 [6] 还将 SSC 扩展到噪声模型, 目标函数如下:

$$\begin{aligned}&\min_z \frac{\lambda}{2}\|X-XZ\|_F^2+\|Z\|_1\\ &\text{s.t.}\quad \text{diag}\{Z\}=0\end{aligned} \tag{3}$$

LSR 最小化 Z 的 Frobenius 范数:

$$\begin{aligned}&\min_z \|Z\|_F^2\\ &\text{s.t.}\quad X=XZ,\quad \text{diag}\{Z\}=0\end{aligned} \tag{4}$$

对于有噪声的模型, 可以扩展为

$$\begin{aligned}&\min_z \|X-XZ\|_F^2+\lambda\|Z\|_F^2\\ &\text{s.t.}\quad \text{diag}\{Z\}=0\end{aligned} \tag{5}$$

去除约束条件, 扩展为

$$\min_z \|X-XZ\|_F^2+\lambda\|Z\|_F^2 \tag{6}$$

其中, $\lambda>0$, $\|Z\|_F$ 是矩阵 Z 的 Frobenius 范数.

2 局部子空间聚类

为解决子空间聚类方法不能直接应用于非线性图像数据的不足, 本文引入 k 近邻局部线性表示思想, 将其分别与经典子空间聚类方法 SSC 和 LSR 相结合, 提出局部子空间聚类 (Local subspace clustering, LSC), 其中 k 近邻局部线性表示与 LSR 结合, 命名为局部最小二乘回归子空间聚类 (Local least squares regression, LLSR); 与 SSC 结合, 命名为局部稀疏子空间聚类 (Local sparse subspace clustering, LSSC).

2.1 局部最小二乘回归子空间聚类

LLSR 算法主要利用 k 近邻[18] 求取 k 个近邻点, 以局部线性为依据对数据进行线性表示. 具体步骤如下:

1) 以样本点间的欧氏距离作为相似性度量, 距离越小表示样本相似性越高. 样本点 x_i 的 k 个近邻点 (不包含自身) 集用 $N_k(x_i)=\{x_{i_1},x_{i_2},\cdots,x_{i_k}\}$ 表示.

模型 (3) 和模型 (5) 中, 对角为 0 的约束本质上是从样本点 x_i 用除自身以外的样本点线性表示得来, 即 $z_{ii}=0$ $(i=1,2,\cdots,n)$, 因此 k 个近邻点中不包含 x_i 自身的目的是为保持模型中对角元素为 0 的约束.

2) LLSR 子空间聚类的目标函数为

$$\min_z \sum_i \|x_i-\sum_{x_j\in N_k(x_i)}x_j z_{ij}\|_F^2+\lambda\|Z\|_F^2 \tag{7}$$

其中若 $x_j\notin N_k(x_i)$, 则令 $z_{ij}=0$. 由于每个 x_i 相互独立, 故上式亦可表示为

$$\min_{\tilde{z}_i} \|x_i-N_k(x_i)\tilde{z}_i\|_F^2+\lambda\|\tilde{z}_i\|_F^2 \tag{8}$$

利用最小二乘思想求解式 (8), 得到解析解为

$$\tilde{z}_i=(N_k(x_i)^{\mathrm{T}}N_k(x_i)+\lambda I)^{-1}N_k(x_i)^{\mathrm{T}}x_i \tag{9}$$

对于每个样本点 x_i, LLSR 都可求解出一个解 $\tilde{z}_i$. 对任意 i, $z^*_{ii_m}=\tilde{z}_{ii_m}$, $m=1,2,\cdots,k$, 其余为 0, 故可得到模型 (7) 的解 Z^*.

2.2 局部稀疏子空间聚类

由于近邻参数 k 值固定, 不能根据样本点分布的疏密自主确定近邻点个数, 导致第 2.1 节提出的 LLSR 方法存在局限, 为此本节提出局部稀疏子空间聚类 LSSC. LSSC 利用 l_1 范数具有稀疏的性质, 在 k 近邻点选取过程中可自适应地调整近邻点个数. 由于 SSC 与 LSR 的正则参数项位置不同, 为了让正则参数项的作用统一, 将 SSC 改写成:

$$\min_z \frac{1}{2}\|X-XZ\|_F^2+\lambda\|Z\|_1 \tag{10}$$

此式可利用交替方向乘子法 (Alternating direction method of multipliers, ADMM)[19] 的思想求解. 本文的 LSSC 在式 (10) 上进行局部线性化改进, 其算法步骤与 LLSR 相同. 先选取每个样本点的 k 个近邻点, 再从近邻点中二次筛选, 自动调整近邻. LSSC

目标函数如下:

$$\min_{z} \frac{1}{2}\sum_{i}\|x_i - \sum_{x_j\in N_k(x_i)} x_j z_{ij}\|_F^2 + \lambda\|Z\|_1 \quad (11)$$

其中, 若 $x_j \notin N_k(x_i)$, 则令 $z_{ij}=0$. 同样, 每个 x_i 相互独立, 故式 (11) 亦可为

$$\min_{\tilde{z}_i} \frac{1}{2}\|x_i - N_k(x_i)\tilde{z}_i\|_F^2 + \lambda\|\tilde{z}_i\|_1 \quad (12)$$

此目标函数与 LASSO 的回归模型等价, 利用 ADMM, 目标函数 (12) 转化为

$$\begin{aligned} &\min_{\tilde{z}_i} \frac{1}{2}\|x_i - N_k(x_i)\tilde{z}_i\|_F^2 + \lambda\|\beta\|_1 \\ &\text{s.t.} \quad \tilde{z}_i - \beta = 0 \end{aligned} \quad (13)$$

其中 β 是对偶变量. 进一步可写成:

$$\begin{aligned} \min_{\tilde{z}_i,\beta,U} &\frac{1}{2}\|x_i - N_k(x_i)\tilde{z}_i\|_F^2 + \lambda\|\beta\|_1 + \\ &\|\frac{\rho}{2}\|\tilde{z}_i - \beta - U\|_F^2 \end{aligned} \quad (14)$$

其中 ρ 是拉格朗日乘子, 在实验中设为常数 1, U 是对偶变量. 模型 (14) 的 ADMM 步骤如下:

$$\begin{aligned} \tilde{z}_i^{K+1} &:= (N_k(x_i)^{\mathrm{T}}N_k(x_i) + \rho I)^{-1}(N_k(x_i)x_i + \\ &\quad \rho\beta^K - U^K) \\ \beta^{K+1} &:= S_{\frac{\lambda}{\rho}}(\tilde{z}_i^{K+1} + \frac{U^K}{\rho}) \\ U^{K+1} &:= U^K + \rho(\tilde{z}_i^{K+1} - \beta^{K+1}) \end{aligned} \quad (15)$$

其中 K 是更新次数, $S_\alpha(v) = (v-\alpha)_+ - (-v-\alpha)_+$, 具体算法参见文献 [19], 由此可求得 $\tilde{z}_i$. 对任意 i, 在相应位置上 $z^*_{ii_m} = \tilde{z}_{ii_m}$, $m = 1, 2, \cdots, k$, 其余为 0, 可得到 LSSC 模型 (11) 的解矩阵 Z^*.

2.3 局部子空间聚类算法

类似于 SSC、LRR 和 LSR, 局部子空间聚类算法 LSC 也是基于谱聚类的子空间聚类方法. LSC 算法如下:

算法 1. 局部子空间聚类算法

输入. 数据矩阵 X, 正则参数 λ, 近邻点个数 k, 类数 c.

输出. c 个类簇.

1) 利用 k 近邻求每个样本点的 k 个近邻;

2) 通过 LLSR 目标函数 (7) 或 LSSC 目标函数 (11) 求解矩阵 Z^*;

3) 计算仿射矩阵 $(|Z^*| + |(Z^*)^{\mathrm{T}}|)/2$;

4) 应用标准分割方法[20] 将数据分成 c 个子空间.

值得注意的是, 文献 [8] 证明了 LSR 具有聚集性, 同样我们可以证明 LLSR 具有局部聚集性.

定理 1. 给定向量 $y \in \mathbf{R}^d$, 矩阵 $N_k(y) \in \mathbf{R}^{d\times k}$ 和参数 $\lambda > 0$, 假设 $N_k(y) = \{y_{(1)}, y_{(2)}, \cdots, y_{(k)}\}$ 已经按列标准化, z^* 是如下 LLSR 问题的解:

$$\min_{z} \|y - N_k(y)z^*\|_2^2 + \lambda\|z\|_2^2 \quad (16)$$

那么

$$\frac{\|z^*_{(i)} - z^*_{(j)}\|_2}{\|y\|_2} \le \frac{1}{\lambda}\sqrt{2(1-r)} \quad (17)$$

其中 $r = y_{(i)}^{\mathrm{T}}y_{(j)}$ 是样本相关系数.

证明. 令 $L(z) = \|y - N_k(y)z^*\|_2^2 + \lambda\|z\|_2^2$, 且

$$\frac{\partial L(z^*)}{\partial z_k} = 0 \quad (18)$$

有:

$$-2y_{(i)}^{\mathrm{T}}(y - N_k(y)z^*) + 2\lambda z^*_{(i)} = 0 \quad (19)$$

$$-2y_{(j)}^{\mathrm{T}}(y - N_k(y)z^*) + 2\lambda z^*_{(j)} = 0 \quad (20)$$

由式 (19) 和式 (20) 得

$$z^*_{(i)} - z^*_{(j)} = \frac{1}{\lambda}(y_{(i)}^{\mathrm{T}} - y_{(j)}^{\mathrm{T}})(y - N_k(y)z^*) \quad (21)$$

且 $N_k(y)$ 已经按列标准化, $r = y_{(i)}^{\mathrm{T}}y_{(j)}$, 得

$$\|y_{(i)} - y_{(j)}\|_2 = \sqrt{2(1-r)} \quad (22)$$

又因为 z^* 是问题 (16) 的最优解, 得

$$\|y - N_k(y)z^*\|_2^2 + \lambda\|z^*\|_2^2 = L(z^*) \le L(0) = \|y\|_2^2 \quad (23)$$

因此, 结合式 (21) 和式 (22) 可得

$$\frac{\|z^*_{(i)} - z^*_{(j)}\|_2}{\|y\|_2} \le \frac{1}{\lambda}\sqrt{2(1-r)} \quad (24)$$

□

定理 1 证明了 LLSR 的局部聚集能力. 假如 $y_{(i)}$ 和 $y_{(j)}$ 是高度相关的, 如 $r = 1$, 则定理 1 表明 $y_{(i)}$ 和 $y_{(j)}$ 对应的系数 $z_{(i)}$ 和 $z_{(j)}$ 的差异性接近为 0, 这样 $y_{(i)}$ 和 $y_{(j)}$ 就会聚成相同的簇.

3 实验

为验证局部子空间聚类方法的有效性, 本节选用经典聚类方法: K 均值 (K-means)、层次聚类方法 (Hierarchical clustering, HC), 以及子空间聚类方法: SSC[6]、LRR[7]、LSR[8]、BD-LRR[12]、RLLRR[9] 和 SMR[10], 与本文提出的 LLSR 和 LSSC 进行比较, 并从聚类准确率 (Accuracy, ACC)[21] 的角度对比各种方法的聚类能力, 其中 ACC 定义如下:

对给定样本, 令 r_i 和 s_i 分别为聚类算法得到的类标签和样本自带的类标签, 准确率计算公式为

$$ACC = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n} \tag{25}$$

其中, n 为样本总数, $\delta(x, y)$ 是一个函数, 当 $x = y$ 时, 值为 1, 否则为 0. $map(r_i)$ 是一置换函数, 将每个类标签 r_i 映射成与样本自带的类标签等价的类标签.

本文采用双月形数据、图像数据集和基因表达数据集, 其中图像数据集包括: ORL10P[1]、PIX10P[1]、PIE10P[1]、Umist[22]、USPS[10]、COIL20[2]; 基因数据集[3]包括: Leukemia1、SRBCT、Lung_Cancer、Prostate_Tumor. 在参数设置方面, LRR、LSR 和 LLSR 的正则参数 λ 取值为 $\{0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10\}$, SSC 和 LSSC 的 λ 取值为 $\{0.0001, 0.001, 0.005, 0.01, 0.05, 0.1\}$; BD-LRR、RLLRR 和 SMR 的参数取 $\{0.08, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1\,000\}$; 近邻点个数 k 取 3～10. 每个算法运行 10 次, 取聚类准确率的平均值. 本文的实验环境为 Winows 7 系统, 内存 4 GB 和 i5 处理器, 所有方法都用 Matlab2012b 编程实现.

3.1 人造数据

利用人造双月形实验数据说明 LSC 的有效性, 同时展示近邻点 k 值和正则参数 λ 的选取对聚类准确率的影响.

双月形数据是由两类数据组成的非线性数据, 形状如两个弯月, 如图 2 所示. 我们的目标是希望能找到一个好的仿射矩阵, 将数据准确地分成两类. 利用 LLSR 和 LSSC 学习得到的邻接图如图 3 所示.

结合图 1 和图 3, 易知 LLSR 和 LSSC 能够很好地将两类数据分开, 便于聚类. 比较图 3 中 LLSR 和 LSSC 的邻接图, 后者以稀疏为目的能够自动筛选近邻点, 故邻接边较少. 表 1 给出各对比方法在双月形数据上的聚类准确率 (参数) 和运行时间. 由表 1 可以看出 LLSR 和 LSSC 很大程度地提升了聚类准确率, 二者均能达到 100 % 的准确率, 而其余方法基本在 50 % 左右. 比较 SSC 和 LSSC 运行时间可以看出, LSSC 利用局部线性的思想明显提升了运行速度. LSSC 和 LLSR 的运行时间包括近邻点计算的耗时.

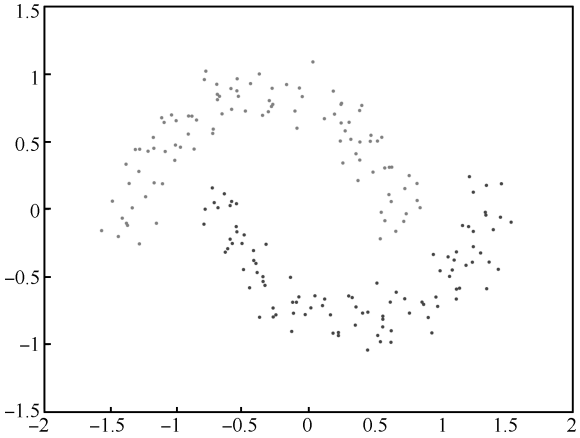
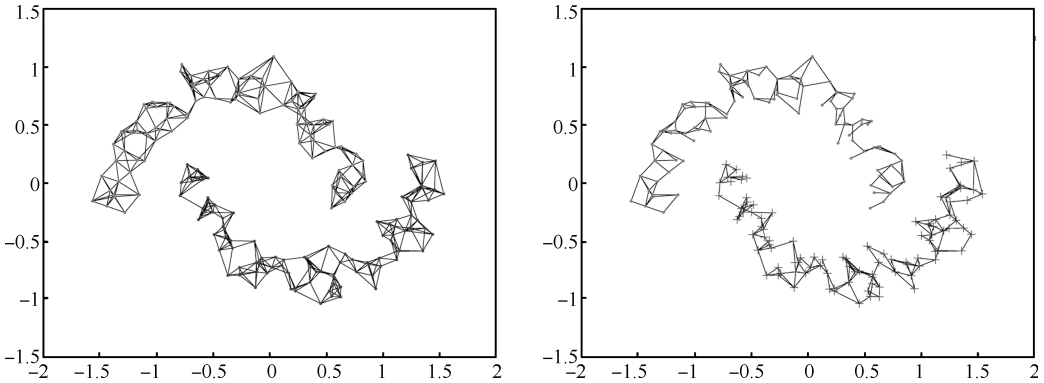


图 2 双月形数据

Fig. 2 The two-moon synthetic data

图 4 给出 LSR、SSC、LLSR 和 LSSC 四种方法的仿射矩阵图. 对于双月形数据, LSR 和 SSC 所



(a) LLSR 学习的邻接图
(a) Learned adjacency graph by LISR

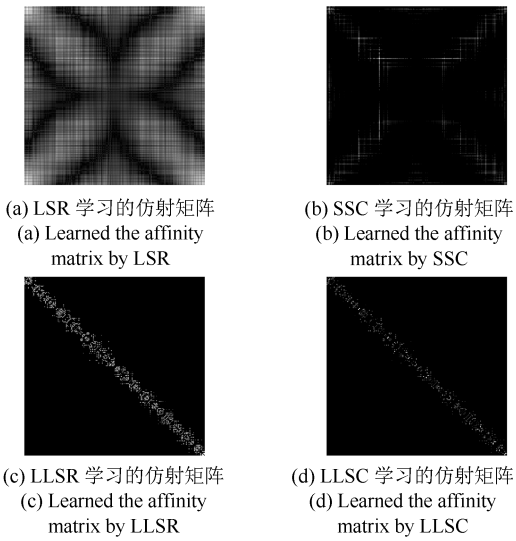
(b) LSSC 学习的邻接图
(b) Learned adjacency graph by LSSC

图 3 LSC 在双月形数据学习得到的邻接图

Fig. 3 Learned adjacency graph by LSC on the two-moon synthetic data

[1]http://featureselection.asu.edu/datasets.php
[2]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[3]http://www.gems-system.org

得的仿射矩阵效果并不理想; 反观, LLSR 和 LSSC 两种方法都能得到对角形式的仿射矩阵, 且 LSSC 能得到更稀疏的仿射矩阵. 综上, 对于双月形数据, 原有 LSR 和 SSC 方法受全局线性限制, 不能得到理想结果, 而局部子空间聚类 LSC 的两种方法利用局部线性思想得到较好的结果, 更好地解决非线性数据的聚类问题.



(a) LSR 学习的仿射矩阵
(a) Learned the affinity matrix by LSR

(b) SSC 学习的仿射矩阵
(b) Learned the affinity matrix by SSC

(c) LLSR 学习的仿射矩阵
(c) Learned the affinity matrix by LLSR

(d) LLSC 学习的仿射矩阵
(d) Learned the affinity matrix by LLSC

图 4　双月形数据上的仿射矩阵

Fig. 4　The affinity matrixes on the two-moon synthetic data

在双月形数据实验中研究正则参数 λ 和近邻点 k 的选取对聚类准确率的影响, 结果如图 5 所示.

由图 5 (a) 可知, 将 LSR 和 SSC 应用在非线性数据上, 正则参数 λ 的改变不会使聚类准确率发生较大变化; 而对于 LLSR, 由于采用局部线性思想, 相当于在很大程度上剔除了许多不相关点, 因此 λ 的变化也不会影响其聚类准确率; LSSC 也是如此, 在大部分 λ 取值下, 得到的结果是稳定的, 但当 λ 取值越大时, l_1 范数的作用越大, 会导致最终获得的近邻点减少, 甚至近邻点个数为 0, 得到的仿射矩阵为 0 矩阵, 影响了聚类准确率. 所以, 图 5 (a) 中 SSC 和 LSSC 在取较大正则参数时, 聚类准确率均发生下降, 甚至在过大的正则参数下无法进行有效聚类. 图 5 (b) 显示近邻点个数的变化会导致聚类准确率的变化, 当 k 取过小或过大, 都会使得聚类准确率降低, 当 k 取 5 ∼ 7 时, LLSR 和 LSSC 会得到较好的结果. LLSR 整体上比 LSSC 更稳定, 有更高的聚类准确率. 在选取近邻点时, LSSC 的 k 值一般比 LLSR 小. 且实验中发现, LSSC 虽然能自动调整近邻点, 但当近邻点数 k 过大时, l_1 范数自动调整近邻点的效果变差.

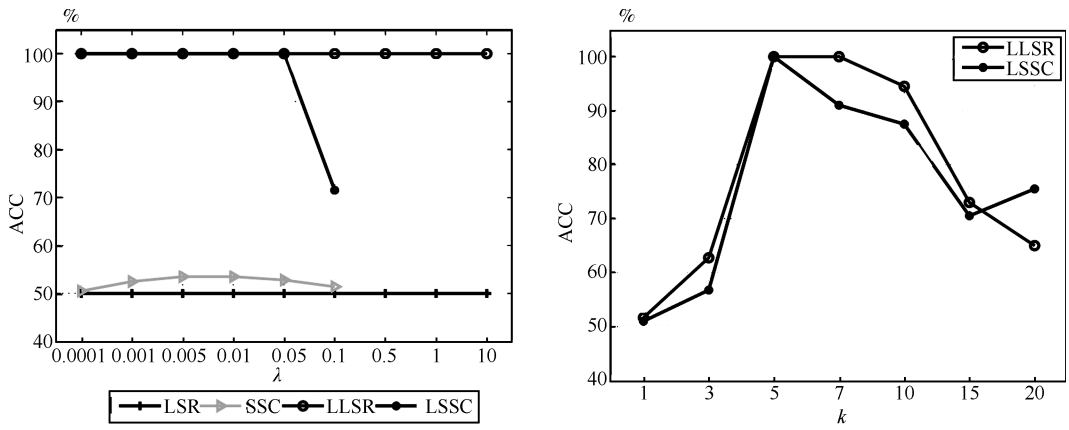
3.2　图像数据

本实验采用的 6 个图像数据中, 前 4 个是人脸

表 1　双月形数据上聚类准确率 (%) 和运行时间 (s) 的对比

Table 1　Clustering accuracy (%) and running time (s) comparison on the two-moon synthetic data

	HC	K-means	LRR	SSC	LSR	BD-LRR	RLLRR	SMR	LSSC	LLSR
ACC	100.00	72.00	53.50 (0.001)	53.50 (0.005)	50.00 (0.0001)	50.00 (0.08)	52.00 (0.1)	51.50 (0.001)	**100.00 (0.0001,5)**	**100.00 (0.0001,5)**
Time	0.0010	0.0026	1.89	4.80	**0.0008**	19.15	0.33	0.045	0.94	0.10



(a) 学习 λ ($k = 5$)
(a) Learned λ ($k = 5$)

(b) 学习 k ($\lambda = 0.05$)
(b) Learned k ($\lambda = 0.05$)

图 5　在双月形数据上 LSC 的参数学习

Fig. 5　Study on the LSC′s parameters on the two-moon synthetic data

表 2 数据集描述

Table 2 Summary of the data sets

数据集	样本	长	宽	类别
ORL10P	100	112	92	10
PIX10P	100	100	100	10
PIE10P	210	55	44	10
Umist	575	28	23	20
USPS	1 000	16	16	10
COIL20	1 440	32	32	20

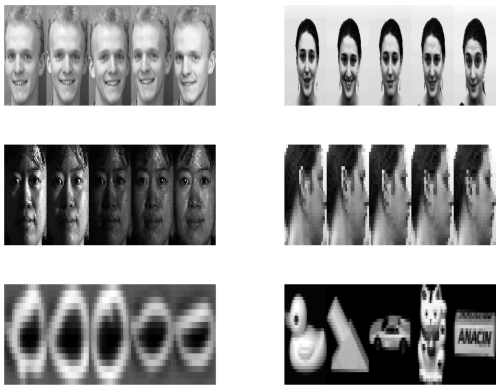


图 6 部分样本图像

Fig. 6 Sample images

图像数据, 第 5 个是手写数字图像数据, 第 6 个为物品数据. 数据集描述如表 2. 且图 6 列出 6 个图像数据集中的部分样本图像.

3.2.1 聚类准确率

实验时, 使用 PCA 统一将 6 个图像数据降至 60 维. 实验结果如表 3.

由表 3 可知, 局部子空间聚类方法 LLSR 和 LSSC 利用局部线性思想, 在所有测试数据集上都取得最好结果. 在大部分图像数据中, LSC 的聚类准确率提高明显. 对比两种不同的局部子空间聚类方法可以发现: LLSR 的聚类准确率除 Umist 数据集外均优于 LSSC, 较好地改进了最小二乘子空间聚类算法. 值得注意的是, 在 PIE10P 实验中, 其他传统子空间聚类方法也取得 90 % 以上较高的聚类准确率, 说明该数据集的线性相关度非常高, 因此 k 值选取较大, 局部近邻点几乎包括全部样本数据. 由此可见 LSC 还具有剔除噪声点的作用. 在其余 5 个数据集中, 近邻点 k 都在 5 ∼ 7 之间. 实验结果表明局部子空间聚类方法能更好地处理具有非线性特征的图像数据, 可以更好地反映图像数据的本质结构.

3.2.2 运行时间

表 4 给出几种对比算法的运行时间, 数据集按样本数由少到多排列. 由表 4 可以看出随着样本数的增大, 各算法的运行时间也相应增加. 几种子空间聚类方法中 LSR 速度最快, 最慢的方法是 SSC. 在大部分实验中, LSC 的两个方法会比 SSC、LRR、BD-LRR 和 RLLRR 快. 特别是由 SSC 改进而来的 LSSC, 仅选取少数近邻点学习仿射矩阵, 其计算量明显减少, 因此运行时间比 SSC 节省很多, 且随着数据集规模的增大, LSSC 较之 SSC 的优势也更明显.

3.2.3 PCA 对聚类准确率的影响

PCA 是一种线性降维方法, 使得数据在降维后

表 3 聚类准确率 (%)

Table 3 Clustering accuracy (%)

	HC	K-means	LRR	SSC	LSR	BD-LRR	RLLRR	SMR	LSSC	LLSR
ORL10P	41.00	73.40	79.00	71.00	83.00	70.30	74.70	78.00	86.00	**87.00**
PIX10P	77.00	79.90	87.00	86.00	85.00	76.80	56.10	88.00	96.00	**97.00**
PIE10P	70.95	32.95	**100.00**	90.00	90.00	80.00	79.43	**100.00**	98.57	**100.00**
Umist	45.57	47.58	52.17	61.57	52.35	48.35	50.96	69.91	**76.87**	74.09
USPS	10.90	73.14	78.60	60.80	71.30	63.90	65.50	77.10	81.20	**91.20**
COIL20	53.47	60.10	65.69	72.01	63.40	67.72	68.80	67.15	78.26	**79.58**

表 4 运行时间的对比 (s)

Table 4 Running time (s) comparison

	HC	K-means	LRR	SSC	LSR	BD-LRR	RLLRR	SMR	LSSC	LLSR
ORL10P	0.0011	0.0071	0.54	0.21	**0.00078**	7.29	1.69	0.014	0.14	0.034
PIX10P	0.00095	0.0062	1.04	0.33	**0.00073**	7.60	1.72	0.012	1.25	0.035
PIE10P	0.0023	0.011	4.51	2.53	**0.0015**	22.23	2.14	0.057	0.32	0.13
Umist	0.011	0.037	25.14	62.27	**0.015**	240.70	13.48	0.71	1.52	0.92
USPS	**0.034**	0.091	130.61	124.57	0.044	884.42	120.33	4.53	3.57	2.75
COIL20	0.072	**0.071**	423.54	2.51	1 446.67	926.78	134.86	18.92	18.97	5.69

既能尽量保持原有数据信息, 又能使算法更快运行, 因此被广泛使用. 由于图像高维性的限制, 若直接采用原始图像数据会因计算机内存不足而导致算法无法执行, 故本实验先对图像数据进行 PCA 降维. 针对本文提出的局部子空间聚类算法, 讨论 PCA 降维对局部子空间聚类算法的影响. 分别将 6 个图像数据集降至 20, 40, 60 和 80 维, 实验结果如图 7.

图 7 中 LSSC 和 LLSR 取固定的近邻数 k ($k = 5$). 由图 7 可知, 在数据为 20, 40, 60 和 80 维时, LLSR 和 LSSC 大部分都能取得最好的结果. SSC 在大部分维数高的情况下, 聚类准确率反而下降, 而 LSSC 与 SSC 相反; LLSR 不随维数的变化而产生大变化, 说明其是一种比较稳定的算法. 图 7 (d) 是侧脸渐变数据, 正脸与侧脸的样本数据差异较大; 用原始 SSC 和 LSR 算法所得的聚类准确率不高, 说明该数据全局线性相关性不强, 而 LSSC 和 LLSR 的聚类准确率会随着维数的增高而增加.

3.3　基因数据

实验时, 用 PCA 统一将 4 个基因表达数据降至 60 维. 4 个基因数据主要信息如表 5 所示.

由表 6 给出的各类算法的聚类准确率可知, 在 4 个基因表达数据集上, 局部子空间聚类方法 LLSR 和 LSSC 都取得最好的聚类效果. 且在大部分数据集上, LLSR 和 LSSC 的聚类准确率比其他方法提高 10 %. 一般情况下参数设置为 k 取 5 到 7. 特别对于 Lung_Cancer 数据, LSR 的聚类准确率达 92 %, 这说明数据全局线性相关, 故选取较大 k 值. 实验结果表明, 局部子空间聚类方法在基因表达数据上也能取得更好的聚类结果.

表 5　数据集描述

Table 5　Summary of the data sets

数据集	样本	基因	类别
Leukemia1	72	5 327	3
SRBCT	83	2 308	4
Lung_Cancer	203	12 600	5
Prostate_Tumor	102	10 509	2

3.4　参数选择

LSC 模型有两个参数: 正则参数和近邻数 k.

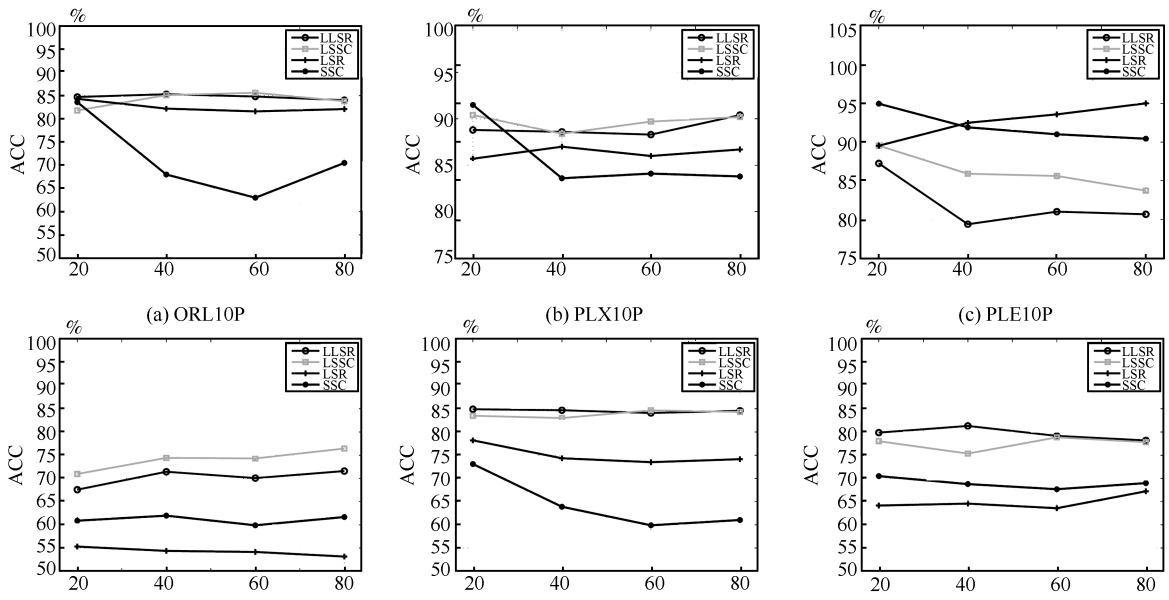


图 7　PCA 对不同图像数据和算法的影响

Fig. 7　PCA on the image data and algorithms

表 6　聚类准确率 (%)

Table 6　Clustering accuracy (%)

	HC	K-means	LRR	SSC	LSR	BD-LRR	RLLRR	SMR	LSSC	LLSR
Leukemia1	54.17	69.31	86.11	58.33	77.78	79.17	54.17	77.78	**90.28**	**90.28**
SRBCT	36.14	53.73	68.43	40.12	54.22	60.24	46.99	63.68	**74.70**	74.46
Lung_Cancer	78.33	83.50	87.39	83.74	**92.61**	85.22	84.24	90.64	91.63	**92.61**
Prostate_Tumor	51.96	63.73	62.75	56.86	62.75	60.78	60.78	59.80	66.67	**69.61**

从图 5 (a) 可以看出, LSC 聚类准确率对参数 λ 的变化不敏感, 主要因为剔除大量不相关样本点, 使得参数 λ 对实验结果的影响减弱, 但还是建议将 λ 选在 $0.001 \sim 0.1$ 之间. 因为 λ 过大, LSSC 求出的仿射矩阵将是 0 矩阵. 另一个参数 k 通常取 $5 \sim 7$ 之间能取得理想结果, 大部分情况下 k 过大或过小可能导致准确率降低; 而对于线性相关程度较大的数据, k 值可取较大值, 这样可起到剔除噪声点的作用. 综上, 对于大部分数据集, λ 取 $0.001 \sim 0.1$ 且 k 取 $5 \sim 7$ 时, LSC 会得到较为理想的聚类结果.

4 结论

考虑到实际应用中很多数据具有非线性特征, 将 k 近邻局部线性表示引入子空间聚类, 分别提出 LSSC 和 LLSR 两种局部子空间聚类算法, 两种算法均能较好克服原有算法全局线性的局限, 较好地应用于非线性数据. 同时, 证明了 LLSR 模型具有局部聚集性, 能够聚集局部样本. 在图像数据和基因表达数据的实验中, 发现 LSSC 和 LLSR 均优于传统子空间聚类方法和其他对比方法, 不仅在聚类准确率上有较大提高, 且在运行时间上 LSSC 也取得明显改进. 研究中发现 LLSR 和 LSSC 的聚类效果受 k 值影响, 如何高效选取近邻数 k 将在以后的研究中给出.

References

1 Yang A Y, Wright J, Ma Y, Sastry S S. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 2008, **110**(2): 212−225

2 Vidal R, Tron R, Hartley R. Multiframe motion segmentation with missing data using power factorization and GPCA. *International Journal of Computer Vision*, 2008, **79**(1): 85−105

3 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(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)

4 Hong W, Wright J, Huang K, Ma Y. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 2006, **15**(12): 3655−3671

5 Vidal R, Favaro P. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*, 2014, **43**: 47−61

6 Elhamifar E, Vidal R. Sparse subspace clustering. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, FL, USA: IEEE, 2009. 2790−2797

7 Liu G C, Lin Z C, Yu Y. Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning (ICML). Haifa, Israel, 2010. 663−670

8 Lu C Y, Min H, Zhao Z Q, Zhu L, Huang D S, Yan S C. Robust and efficient subspace segmentation via least squares regression. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 347−360

9 Zhang H Y, Lin Z C, Zhang C, Cao J B. Robust latent low rank representation for subspace clustering. *Neurocomputing*, 2014, **145**: 369−373

10 Hu H, Lin Z C, Feng J J, Zhou J. Smooth representation clustering. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 3834−3841

11 Soltanolkotabi M, Elhamifar E, Candés E J. Robust subspace clustering. *The Annals of Statistics*, 2014, **42**(2): 669−699

12 Feng J S, Lin Z C, Xu H, Yan S C. Robust subspace segmentation with block-diagonal prior. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 3818−3825

13 Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, **58**(1): 267−288

14 Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 1970, **12**(1): 55−67

15 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 2001, **96**(456): 1348−1360

16 Lee S R, Heo G S, Lee C Y. Representation and symbolization of motion captured human action by locality preserving projections. *Applied Mathematics & Information Sciences*, 2014, **8**(1): 441−446

17 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, **290**(5500): 2323−2326

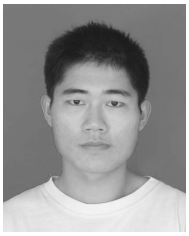
18 Tang Y Y, Yuan H L, Li L Q. Manifold-based sparse representation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, **52**(12): 7606−7618

19 Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011, **3**(1): 1−122

20 Shi J B, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 888−905

21 Cai D, He X F, Wu X Y, Han J W. Non-negative matrix factorization on manifold. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). Pisa: IEEE, 2008. 63−72

22 Hou C P, Nie F P, Yi D Y, Tao D C. Discriminative embedded clustering: a framework for grouping high-dimensional data. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(6): 1287−1299



刘展杰 福州大学数学与计算机科学学院硕士研究生. 主要研究方向为数据挖掘, 模式识别.
E-mail: liufzu@gmail.com
(**LIU Zhan-Jie** Master student at the College of Mathematics and Computer Science, Fuzhou University. His research interest covers data mining and pattern recognition.)

陈晓云 福州大学数学与计算机科学学院教授. 主要研究方向为数据挖掘、模式识别. 本文通信作者.
E-mail: c_xiaoyun@21cn.com
(**CHEN Xiao-Yun** Professor at the College of Mathematics and Computer Science, Fuzhou University. Her research interest covers data mining and pattern recognition. Corresponding author of this paper.)