文章编号: 1001-0920 (2016) 07-1272-07 DOI: 10.13195/j.kzyjc.2015.0649

基于流形鉴别信息的特征选择及其结构化稀疏表示

钱 彬, 唐振民, 李 雪, 徐 威

(南京理工大学 计算机科学与工程学院，南京 210094)

摘 要: 针对启发式特征选择策略忽略了特征间相关信息导致子最优的问题, 提出一种基于流形鉴别信息的特征选择 (MDFS) 算法. 该算法根据近邻信息和标签信息刻画高维数据类内和类间流形结构, 以最小化流形散度差为准则构建目标函数, 并增加结构化稀疏正则项降低特征间冗余. 通过统一框架下的特征权重迭代优化获得最优特征子集. 在 ORL 库、COIL20 库、Isolet1 库上的聚类实验表明, MDFS 算法选取的特征子集相比传统算法具有更高的识别准确率和归一化互信息, 验证了所提出算法的有效性.

关键词: 特征选择；流形学习；结构化稀疏；聚类

中图分类号: TP391 **文献标志码:** A

Feature selection based on manifold discriminant information and its structured sparse representation

QIAN Bin, *TANG Zhen-min*, *LI Xue*, *XU Wei*

(School of Computer Science and Engineering，Nanjing University of Science and Technology，Nanjing 210094，China. Correspondent：QIAN Bin，E-mail：311062198@njust.edu.cn)

Abstract: The traditional heuristic feature selection methods usually neglect the correlations between features, and thus lead to suboptimal feature subset. Therefore, a method of manifold discriminant feature selection(MDFS) is proposed. The method captures the manifold structure of the dataset by incorporating both neighbor and label information, and then the objective function can be formulated by minimizing the difference between intra and inter scatters. Besides, the structured sparse regularization term is further added to reduce the redundant information. Finally, a new iterative algorithm is presented for optimization. The experimental results on three popular datasets, i.e., ORL, COIL20, and Isolet1 dataset, show that, compared with existing related methods, the proposed method achieves better clustering performances in terms of accuracy and normalized mutual information. Thus the effectiveness of the proposed method can be verified.

Keywords: feature selection；manifold learning；structured sparsity；cluster

0 引 言

近年来, 随着大数据技术的日益发展, 数据降维方法在计算机视觉、模式识别、生物基因技术等方面起着越来越重要的作用[1]. 数据降维方法可以分为两类: 特征抽取和特征选择. 特征抽取通过某些准则寻找高维数据的低维子空间进行投影降维, 而特征选择在不改变原始特征数据的基础上对特征进行筛选, 去除冗余数据, 保留对于分类或识别具有重要意义的特征. 因为特征选择在数据降维的同时没有改变特征的物理意义, 所以基于特征选择的降维方法已被广泛应用于各个领域[2-5].

根据特征选择方法是否依赖于最终采用的学习分类器, 可以分为两种类型: 封装式 (Wrapper) 和过滤式 (Filter)[6]. Guyon 等[7]将支持向量机分类器应用于基因特征选择; Michalak 等[8]提出了一种基于相关性的封装式特征选择方法, 这类封装式的特征选择算法需要依赖特定的分类器, 其计算复杂, 算法适应性较差. 基于方差 (VAR) 的特征评判准则由于其计算简单受到广泛使用, 但是没有考虑样本之间的依赖关系, 因而所选取的特征不具有良好的表示能力. He 等[9]提出了拉普拉斯排序 (LS) 算法, 该算法通过构建样本拉普拉斯近邻图, 以特征局部保持能力为准则对样

收稿日期: 2015-05-24；**修回日期:** 2015-12-12.

基金项目: 国家自然科学基金项目(61305134, 90820306)；江苏省社会安全图像与视频理解重点实验室基金项目(30920130122006).

作者简介: 钱彬(1989−), 男, 博士生, 从事计算机视觉、模式识别的研究；唐振民(1961−), 男, 教授, 博士生导师, 从事智能机器人与目标识别、图像处理与模式识别等研究.

本特征权重进行排序, 采用启发式策略逐个选取最优特征构成特征子集, 但是这种方式没有考虑到特征之间的相关性, 得到的特征子集并不能保证子集最优. Nie 等[10]提出了一种基于迹比准则 (TRC) 的特征选择方法, 通过构建类内和类间散度, 以迭代方式更新特征权重, 可以一次性获得最优特征子集, 但是该方法没有对特征权重进行有效约束, 所选取的特征存在大量冗余信息. Cai 等[11]提出了一种多聚类特征选择 (MCFS) 算法, 在谱回归 (SR) 的基础上对特征权重加以 L_1 范数约束, 使得特征呈现有效的稀疏化特性, 提高了特征局部保持能力, 有效减少了特征冗余信息, 但是该算法在实现时需要对多个特征向量进行谱回归, 容易丢失有效鉴别信息, 并且最后对多个特征权重向量融合的方法不能保证全局最优.

为了有效利用样本标签以及保持样本间邻域结构关系, 本文提出了一种基于流形鉴别信息的特征选择算法, 根据样本近邻信息和标签信息刻画类内和类间流形结构, 以最小化流形散度差为准则, 使得选取的特征具有良好的鉴别能力和流形结构保持能力, 同时对样本特征权重矩阵采用 $L_{2,1}$ 范数[12-13]进行结构化稀疏约束, 进一步减少数据间的冗余信息以提高特征表示能力. 本文引入中间辅助变量采用梯度下降算法对目标函数进行迭代优化, 针对样本个数和样本维数的大小关系, 采用直接计算或谱嵌入[11]两种模式进行求解.

1　相关工作

1.1　范数正则化

Cai 等[11]在谱回归的基础上增加 L_1 范数正则化, 使得选取的特征权重向量趋于 0, 仅保留个别非 0 元素作为最终选取的特征, 有效地减少了特征间冗余信息, 并在降维的同时形成了特征向量良好的稀疏表示. 对于任意向量 $\boldsymbol{v} \in \boldsymbol{R}^d$, 其 L_p 范数定义如下:

$$\|\boldsymbol{v}\|_p = \Big(\sum_{i=1}^{d} |v_i|^p\Big)^{\frac{1}{p}}. \tag{1}$$

L_p 范数约束属于向量约束, 而常规的特征选择方法往往将特征权重定义为投影矩阵形式, 通过寻找最优鉴别子空间对矩阵进行权重更新, 这需要对矩阵进行范数正则化. 通过谱回归可以规避这一问题, 但是需要对多个回归向量进行特征融合, 其融合算法的优劣会影响最终特征选择的结果. Nie 等[13]采用了 $L_{2,1}$ 范数对矩阵进行约束, 并给出了有效的收敛性证明. 对于任意矩阵 $\boldsymbol{M} \in \boldsymbol{R}^{d\times n}$, m^i 表示 $\boldsymbol{M}$ 的第 i 行, m_j 表示 $\boldsymbol{M}$ 的第 j 列. 矩阵 $\boldsymbol{M}$ 的 $L_{2,1}$ 范数定义如下:

$$\|\boldsymbol{M}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{n} m_{ij}^2} = \sum_{j=1}^{d} \|m^i\|_2. \tag{2}$$

从式 (2) 的定义可以看出, $L_{2,1}$ 范数对每列数据进行 L_2 范数的平方约束, 对于每行数据进行 L_1 范数约束, 形成了行数据之间的竞争, 使得行数据趋于 0, 仅保留个别非 0 行数据, 从而实现对于矩阵的结构化稀疏约束.

1.2　拉普拉斯特征排序 (LS)

He 等[9]提出了拉普拉斯特征排序 (LS) 算法, 以保持高维样本局部近邻关系为衡量特征权重的准则, 采用无监督方式逐个选取最优特征组成特征子集. 定义第 r 个特征的特征权重为 l_r, $f_{r,i}$ 表示第 i 个样本的第 r 个特征, 其中 $i = 1, 2, \cdots, n$. 对于第 r 个特征, 定义其特征向量 $f_r = [f_{r,1}, f_{r,2}, \cdots, f_{r,n}]^{\mathrm{T}}$, 算法步骤如下.

Step 1: 构建 n 个样本节点的近邻图 G. 如果样本 x_i 属于样本 x_j 的近邻或者样本 x_j 属于样本 x_i 的近邻, 则连接节点 i 和 j, 连接边权重设置为

$$\boldsymbol{S}_{i,j} = \mathrm{e}^{-\frac{\|x_i - x_j\|^2}{t}}.$$

Step 2: 定义全 1 向量 $\mathbf{1} = [1, 1, \cdots, 1]^{\mathrm{T}}$ 和对角矩阵 $\boldsymbol{D}$, 对角线元素为 $\boldsymbol{D}_{ii} = \sum_{j=1}^{n} S_{ij}$, 即 $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{S}_{\mathbf{1}})$, 对应的拉普拉斯矩阵为 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$. 对原始特征进行平均值移除, 可以得到

$$\tilde{f}_r = f_r - \frac{f_r^{\mathrm{T}} \boldsymbol{D}_{\mathbf{1}}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{D}_{\mathbf{1}}} \mathbf{1}. \tag{3}$$

Step 3: 计算第 r 个特征的特征权重为

$$l_r = \frac{\tilde{f}_r^{\mathrm{T}} \boldsymbol{L} \tilde{f}_r}{\tilde{f}_r^{\mathrm{T}} \boldsymbol{D} \tilde{f}_r}. \tag{4}$$

最后 LS 算法采用启发式策略逐个选取权重最大的特征作为特征子集. LS 算法有效利用了高维样本流形结构信息, 算法复杂度低, 计算简单, 但是仍存在两个不足: 1) 在流形结构内没有有效利用样本标签信息, 所选取的特征并不一定具有最佳的分类鉴别能力; 2) 由于样本特征之间存在相关性, 采用启发式策略逐个选取最优特征不能保证选取的特征子集具有最优的分类性能.

2　基于流形鉴别的特征选择算法 (MDFS)

2.1　MDFS 算法原理

为了有效利用高维流形中的样本标签信息, 本文提出一种基于流形鉴别的特征选择算法, 通过构建样本类内和类间近邻图, 以流形散度差为根本准则, 在低维空间内寻找类内最小、类间最大的鉴别投影矩阵, 同时对投影矩阵进行 $L_{2,1}$ 范数约束, 使得特征之间相互竞争, 消除特征冗余, 一次性获得最优特征子集. 针对样本维数和样本个数的大小关系, 在统一的迭代优化框架下提出两种解决方案, 以避免大规模矩阵特征值分解的计算难度. MDFS 算法主要包含两部

分: 融合流形鉴别信息的目标函数构建和目标函数求解算法.

2.2 融合流形鉴别信息的目标函数构建

假设样本集 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\boldsymbol{R}^{d\times n}$, d 表示样本维度, n 表示样本个数; 投影矩阵也即特征权重矩阵为 $\boldsymbol{W}\in\boldsymbol{R}^{d\times r}$; 样本集 $\boldsymbol{X}$ 投影后对应的低维流形样本集为 $\boldsymbol{Y}=[y_1^{\mathrm{T}},y_2^{\mathrm{T}},\cdots,y_n^{\mathrm{T}}]\in\boldsymbol{R}^{n\times r}$, 其中 r 表示投影后的维数, 通常 $r<d$. LS 算法构造了样本近邻图作为高维流形信息, 为了能够融合样本标签信息, 这里分别构造样本类内和类间近邻图. 定义类内近邻矩阵为

$$\boldsymbol{S}_{ij}^{w}=\begin{cases}\mathrm{e}^{-\frac{\|x_i-x_j\|^2}{t}},\ x_i\text{和}x_j\text{属于同类且}x_i\in N_k(j);\\ 0,\ \text{else.}\end{cases}\tag{5}$$

其中 $x_i\in N_k(j)$ 表示样本 x_i 为样本 x_j 的 k 近邻或者样本 x_j 为样本 x_i 的 k 近邻, 对应的类内对角矩阵为 $\boldsymbol{D}^w$, 类内拉普拉斯矩阵为 $\boldsymbol{L}^w$. 同理, 定义类间近邻矩阵为

$$\boldsymbol{S}_{ij}^{b}=\begin{cases}\mathrm{e}^{-\frac{\|x_i-x_j\|^2}{t}},\ x_i\text{和}x_j\text{不属于同类且}x_i\in N_k(j);\\ 0,\ \text{else.}\end{cases}\tag{6}$$

对应的类间对角矩阵为 $\boldsymbol{D}^b$, 类间拉普拉斯矩阵为 $\boldsymbol{L}^b$. 为使投影后的样本类内散度最小, 类间散度最大, 定义流形鉴别散度差为

$$\begin{aligned}&J_{MD}=\\&\frac{1}{2}\sum_{i,j=1}^{n}\|y_i-y_j\|^2\boldsymbol{S}_{ij}^{w}-\frac{1}{2}\sum_{i,j=1}^{n}\|y_i-y_j\|^2\boldsymbol{S}_{ij}^{b}=\\&\frac{1}{2}\sum_{i,j=1}^{n}\|\boldsymbol{W}^{\mathrm{T}}x_i-\boldsymbol{W}^{\mathrm{T}}x_j\|^2\boldsymbol{S}_{ij}^{w}-\\&\frac{1}{2}\sum_{i,j=1}^{n}\|\boldsymbol{W}^{\mathrm{T}}x_i-\boldsymbol{W}^{\mathrm{T}}x_j\|^2\boldsymbol{S}_{ij}^{b}=\\&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}^w-\boldsymbol{S}^w)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}-\\&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}^b-\boldsymbol{S}^b)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}=\\&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}-\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}.\end{aligned}\tag{7}$$

为了避免尺度化因素影响, 对特征权重矩阵增加正交约束, 即 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}$. 从而融合流形鉴别散度差的目标函数为

$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\}-\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}.\tag{8}$$

为了突出特征选择结果、减少特征冗余信息, 对特征权重矩阵 $\boldsymbol{W}$ 增加 $L_{2,1}$ 范数约束. 因此, MDFS 算法最终的目标函数为

$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}-\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}+\varepsilon\|\boldsymbol{W}\|_{2,1},\tag{9}$$

其中 ε 表示稀疏约束参数.

2.3 MDFS 目标函数求解

由于 $L_{2,1}$ 范数约束的影响, MDFS 算法目标函数 (9) 并非平滑可导, 目标函数极值无法直接获得. 本文采用 Nie 等[12]提出的一种引入中间变量迭代求解 $L_{2,1}$ 范数约束的方法. 式 (9) 对应的拉格朗日函数为

$$\begin{aligned}&J(\boldsymbol{W})=\\&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}-\mathrm{Tr}\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}+\\&\varepsilon\|\boldsymbol{W}\|_{2,1}-\mathit{\Lambda}\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}-\boldsymbol{I}\}.\end{aligned}\tag{10}$$

将式 (10) 对 $\boldsymbol{W}$ 求导, 并将导数置为 0, 可以得到

$$\begin{aligned}\frac{\partial J(\boldsymbol{W})}{\partial\boldsymbol{W}}&=2\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}-2\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}+\\&2\varepsilon\boldsymbol{U}\boldsymbol{W}-2\boldsymbol{\mathit{\Lambda}}\boldsymbol{W}=0,\end{aligned}\tag{11}$$

其中中间变量矩阵 $\boldsymbol{U}$ 为对角矩阵, 对角线元素

$$u_{ii}=1/2\|w^i\|_2.\tag{12}$$

式 (12) 可以简化为

$$(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{x}^{\mathrm{T}}+\varepsilon\boldsymbol{U})\boldsymbol{W}=\boldsymbol{\mathit{\Lambda}}\boldsymbol{W}.\tag{13}$$

从式 (13) 可以发现, 当样本维度较小, 也即当 $d<n$ 时, 如果矩阵 $\boldsymbol{U}$ 已知, 则对矩阵 $(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U})$ 进行特征值分解可以获得 $\boldsymbol{W}$, 而当 $\boldsymbol{W}$ 已知时又可以根据式 (12) 求得 $\boldsymbol{U}$. 因此本文采用这种迭代求取方式进行特征权重优化.

当样本维度较高, $d>n$ 时, 会产生高维小样本问题, 对式 (13) 中的矩阵进行特征值分解较为复杂且耗时. 因此, 本文对目标函数进行相应变换, 通过谱回归逆向求解. 如果排除 $L_{2,1}$ 范数约束项, 则式 (9) 的求解方式等价于求解广义特征方程

$$\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{\mathit{\Lambda}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\tag{14}$$

的特征值. 该方程可以通过经典的谱回归理论[11]进行变换求解.

定理 1 假设矩阵 $\boldsymbol{Y}\in\boldsymbol{R}^{n\times r}$ 的每一列为特征方程 $\boldsymbol{W}y=\lambda\boldsymbol{D}y$ 对应的特征向量. 如果存在矩阵 $\boldsymbol{S}\in\boldsymbol{R}^{d\times r}$ 使得 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}=\boldsymbol{Y}$, 则矩阵 $\boldsymbol{S}$ 的每一列为特征方程 $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}s=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}s$ 的相同特征值 λ 对应的特征向量.

证明 根据定义有

$$\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}s=\boldsymbol{X}\boldsymbol{W}y=\boldsymbol{X}\lambda\boldsymbol{D}y=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}s.$$

定理成立. □

根据定理 1, 可以将式 (10) 的特征方程求解任务拆分成两步进行:

1) 首先计算特征方程

$$\boldsymbol{L}^w\boldsymbol{Y}=\boldsymbol{\mathit{\Lambda}}\boldsymbol{L}^b\boldsymbol{Y}\tag{15}$$

的特征向量, 得到低维投影样本 $\boldsymbol{Y}$;

2) 找到 $\boldsymbol{W}$, 使得 $\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2$ 尽量小.

当得到低维投影样本$\boldsymbol{Y}$后, 可以再增加特征权重矩阵的$L_{2,1}$范数实现结构化稀疏约束, 因此MDFS的目标函数(9)可以转换为

$$\min \|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\|\boldsymbol{W}\|_{2,1}. \tag{16}$$

其求解过程与式(11)类似, 对$\boldsymbol{W}$进行求导并置为0后得到

$$\boldsymbol{W}=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}. \tag{17}$$

从式(13)可看出, 依然可以通过迭代方式求解$\boldsymbol{U}$和$\boldsymbol{W}$, 使得目标函数逐步下降.

总结本文提出的MDFS算法具体实现如下.

输入: 样本集$\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\boldsymbol{R}^{d\times n}$, 稀疏化参数$\varepsilon$;

输出: 排序过后的特征权重向量$f\in\boldsymbol{R}^d$.

Step 1: 迭代次数t初始化为0, 中间变量$\boldsymbol{U}$初始值设为单位矩阵, 即$\boldsymbol{U}_0=\boldsymbol{I}\in\boldsymbol{R}^{d\times d}$.

Step 2: 根据式(5)和(6)构造类内和类间近邻矩阵$\boldsymbol{S}^w$、$\boldsymbol{S}^b$, 并计算其对应的拉普拉斯矩阵$\boldsymbol{L}^w$、$\boldsymbol{L}^b$. 如果$d>n$, 则根据式(15)计算低维投影样本$\boldsymbol{Y}$.

Step 3: 如果$d>n$, 则计算特征方程

$$(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U}_t)\boldsymbol{W}=\boldsymbol{\Lambda}\boldsymbol{W}$$

对应的特征向量, 得到$\boldsymbol{W}_{t+1}$; 如果$d\leqslant n$, 则按照式(17)$\boldsymbol{W}_{t+1}=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$计算得到$\boldsymbol{W}_{t+1}$.

Step 4: 根据当前$\boldsymbol{W}_{t+1}$和式(12)计算$\boldsymbol{U}_{t+1}$, 令$t=t+1$. 重复执行Step 3和Step 4直至收敛.

Step 5: 根据$\|w_t^i\|_2$的大小对特征权重向量f进行排序.

2.4 MDFS算法收敛性证明

上一节求解了MDFS算法目标函数的更新规则, 本节将证明MDFS算法在更新规则下收敛. 为了证明MDFS收敛, 从目标函数(9)考虑. 由于$\boldsymbol{L}^w$和$\boldsymbol{L}^b$为对称矩阵, 可以得到

$$\begin{aligned}&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}-\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}=\\&\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{L}^w-\boldsymbol{L}^b)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\}\geqslant 0.\end{aligned}$$

因此, 可以看出目标函数(9)拥有一个松弛下界0, 接下来需要证明MDFS在更新规则下非递增即可证明算法收敛.

定理2　目标函数(9)在更新规则(12)和(13)下是非增函数.

证明　根据定义有: 由线性鉴别分析理论和式(13)可以得到

$$\begin{aligned}\boldsymbol{W}_t=\arg\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\\&\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U}_t)\boldsymbol{W}_{t-1}\}.\end{aligned} \tag{18}$$

从而有

$$\begin{aligned}&\mathrm{Tr}\{\boldsymbol{X}_t^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U}_t)\boldsymbol{W}_t\}\leqslant\\&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}+\varepsilon\boldsymbol{U}_t)\boldsymbol{W}_{t-1}\}\Rightarrow\\&\mathrm{Tr}\{\boldsymbol{W}_t^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}})\boldsymbol{W}_t\}+\\&\varepsilon\sum_{i=1}^d\frac{\|w_t^i\|_2^2}{2\|w_{t-1}^i\|_2}\leqslant\\&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}})\boldsymbol{W}_{t-1}\}+\\&\varepsilon\sum_{i=1}^d\frac{\|w_{t-1}^i\|_2^2}{2\|w_{t-1}^i\|_2}\Rightarrow\\&\mathrm{Tr}\{\boldsymbol{W}_t^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}})\boldsymbol{W}_t\}+\varepsilon\sum_{i=1}^d\|w_t^i\|_2-\\&\varepsilon\Big\{\sum_{i=1}^d\|w_t^i\|_2-\sum_{i=1}^d\frac{\|w_t^i\|_2^2}{2\|w_{t-1}^i\|_2}\Big\}\leqslant\\&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}})\boldsymbol{W}_{t-1}\}+\\&\varepsilon\sum_{i=1}^d\|w_{t-1}^i\|_2-\varepsilon\Big\{\sum_{i=1}^d\|w_t^i\|_2-\sum_{i=1}^d\frac{\|w_{t-1}^i\|_2^2}{2\|w_{t-1}^i\|_2}\Big\}.\end{aligned}$$

根据文献[12], 如下式成立:

$$\begin{aligned}&\sum_{i=1}^d\|w_t^i\|_2-\sum_{i=1}^d\frac{\|w_t^i\|_2^2}{2\|w_{t-1}^i\|_2}\leqslant\\&\sum_{i=1}^d\|w_{t-1}^i\|_2-\sum_{i=1}^d\frac{\|w_{t-1}^i\|_2^2}{2\|w_{t-1}^i\|_2},\end{aligned} \tag{19}$$

且根据式(2)有

$$\sum_{i=1}^d\|w_t^i\|_2=\|W\|_{2,1}, \tag{20}$$

即矩阵$\boldsymbol{W}$的$L_{2,1}$范数为各行向量L_2范数之和, 则可以得到

$$\begin{aligned}&\mathrm{Tr}\{\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_t\}-\\&\mathrm{Tr}\{\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_t\}+\varepsilon\|\boldsymbol{W}_t\|_{2,1}\leqslant\\&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{t-1}\}-\\&\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{t-1}\}+\varepsilon\|\boldsymbol{W}_{t-1}\|_{2,1}.\end{aligned} \tag{21}$$

定理成立. □

定理3　目标函数(16)在更新规则(12)和(17)下是非增函数.

证明　当$\boldsymbol{U}_t$已知时, 易知$\boldsymbol{W}_t$为

$$\min\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\mathrm{Tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}_t\boldsymbol{W}\} \tag{22}$$

的最优解, 即

$$\boldsymbol{W}_t=\arg\min\|\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{U}_t\boldsymbol{W}_{t-1}\}. \tag{23}$$

从而有

$$\begin{aligned}&\|\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\mathrm{Tr}\{\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{U}_t\boldsymbol{W}_{t-1}\}\leqslant\\&\|\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\mathrm{Tr}\{\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{U}_t\boldsymbol{W}_{t-1}\}\Rightarrow\\&\|\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\sum_{i=1}^d\frac{\|w_t^i\|_2^2}{2\|w_{t-1}^i\|_2}\leqslant\\&\|\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\sum_{i=1}^d\frac{\|w_{t-1}^t\|_2^2}{2\|w_{t-1}^i\|_2}\Rightarrow\\&\|\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\sum_{i=1}^d\|w_t^i\|_2-\end{aligned}$$

$$\varepsilon\Big\{\sum_{i=1}^{d}\|w_t^i\|_2-\sum_{i=1}^{d}\frac{\|w_t^i\|_2^2}{2\|w_{t-1}^i\|_2}\Big\}\leqslant$$
$$\|\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\sum_{i=1}^{d}\|w_{t-1}^i\|_2-$$
$$\varepsilon\Big\{\sum_{i=1}^{d}\|w_{t-1}^i\|_2-\sum_{i=1}^{d}\frac{\|w_{t-1}^i\|_2^2}{2\|w_{t-1}^i\|_2}\Big\}.$$

同理, 由式 (2) 和 (9) 可以得到

$$\|\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\|\boldsymbol{W}_t\|_{2,1}\leqslant$$
$$\|\boldsymbol{W}_{t-1}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\varepsilon\|\boldsymbol{W}_{t-1}\|_{2,1}. \tag{24}$$

定理成立. □

由定理 1 和定理 2 可知, MDFS 算法的目标函数在更新规则下是非增的, 又因为目标函数拥有松弛下界 0, 因此 MDFS 算法收敛.

3 实验结果与分析

特征选择是模式识别领域的重要研究内容之一, 为了有效验证本文所提方法的有效性和普适性, 需要采用模式识别领域的常用数据库进行验证, 而图像和语音一直是模式识别领域的重点研究对象, 因此采用 3 个公开的权威数据库进行算法分析来进行综合验证. 本节重点评估 MDFS 算法在 ORL 库、COIL20 库、Isolet1 库上的实验效果, 并与 VAR 算法、LS 算法、TRC 算法、MCFS 算法在相同数据库上的结果进行比较, 统一使用 K 均值聚类算法对所选择的特征进行准确率 (AC) 和归一化互信息 (NMI) 评价[11].

3.1 数据集介绍

本文在人脸库 ORL、图像库 COIL20 和语音库 Isolet1 上进行实验.

1) ORL: 由 40 个人的人脸图像组成, 每人 10 幅, 每幅图像光照、面部表情 (睁眼或闭眼, 笑或不笑)、面部细节 (戴眼镜或不戴眼镜) 不相同. 将 400 幅图像归一化为 32 × 32, 样本个数小于特征维数, 如图 1(a) 所示;

2) COIL20: 由 20 个真实物体组成, 每个物体包含 72 幅不同角度拍摄的图像, 图像大小 32 × 32, 样本个数大于特征维数, 如图 1(b) 所示;

3) Isolet1: 30 人对 26 个英文字母各读两遍, 每次获取 617 维的语音特征作为字母样本, 样本个数大于特征维数.

ORL 库、COIL20 库和 Isolet1 库的具体数据描述如表 1 所示.

表 1 数据描述

数据集	样本数	属性数	类别数
ORL	400	1 024	40
COIL20	1 440	1 024	20
Isolet1	1 560	617	26

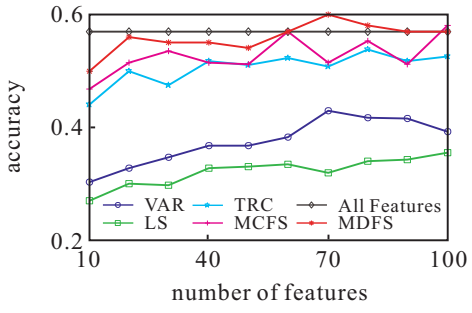
(a) 部分 ORL 人脸库图像

(b) 部分 COIL20 库图像

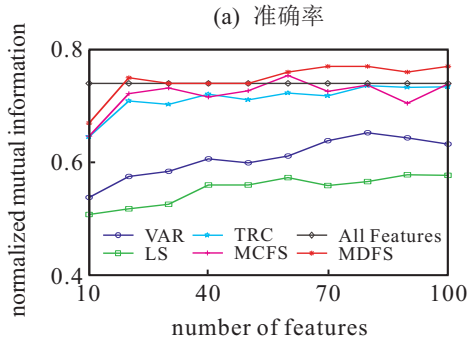
图 1 数据集示例

3.2 实验结果

把数据集提供的类别标签和使用聚类算法得到的类别标签比较后进行算法评估. 图 2 ~ 图 4 分别描述了 5 种算法在 ORL 库、COIL20 库、Isolet1 库上聚类的准确率和归一化互信息. 为了验证本文算法在数据降维上的有效性, 与采用所有特征 (All Features) 同时进行实验对比. 横坐标表示所选特征数, 取前 100 个特征组成最优特征子集对算法进行验证.

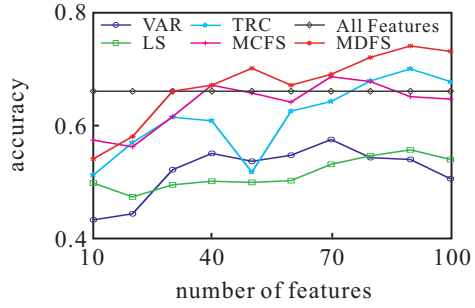


(a) 准确率

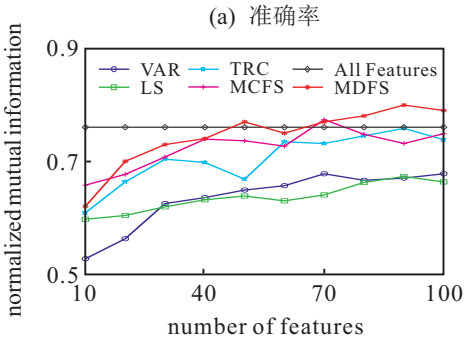


(b) 归一化互信息

图 2 在 ORL 库上聚类的 AC 和 NMI

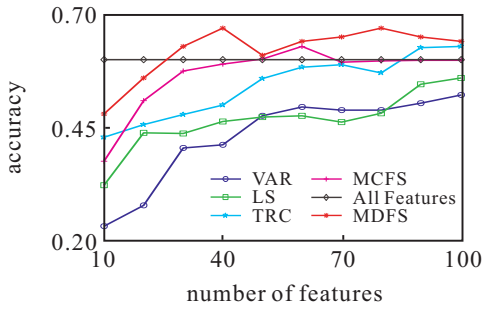


(a) 准确率

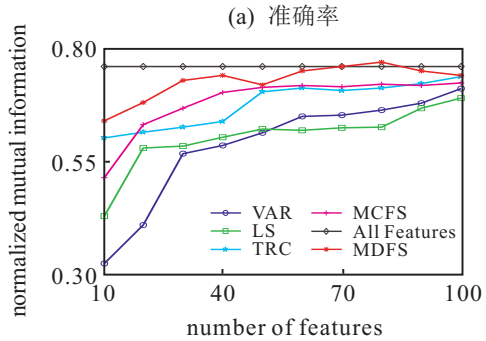


(b) 归一化互信息

图 3　在 COIL20 库上聚类的 AC 和 NMI



(a) 准确率



(b) 归一化互信息

图 4　在 Isolet1 库上聚类的 AC 和 NMI

从图 2～图 4 可以看出, MDFS 算法的平均 AC 和 NMI 高于其他几种算法. 即使采用前 100 个特征, 通过本文提出的 MDFS 算法也可以达到与采用 All Features 的相似或更优越的效果. 表 2 和表 3 详细列出了各算法在数据库上的最高 AC 和 NMI 的大小.

表 2　聚类实验 AC 对比　　%

数据库	VAR	LS	TRC	MCFS	All Features	MDFS
ORL	43.2	35.5	53.7	58.1	57.2	**60.1**
COIL20	57.4	55.6	69.9	68.5	66.3	**74.2**
Isolet1	52.1	55.9	62.9	62.8	60.1	**67.1**

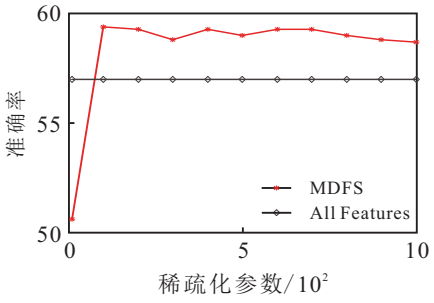
表 3　聚类实验 NMI 对比　　%

数据库	VAR	LS	TRC	MCFS	All Features	MDFS
ORL	65.3	57.7	73.6	75.4	74.3	**77.3**
COIL20	67.8	67.3	75.8	77.3	76.0	**80.1**
Isolet1	71.1	69.0	73.7	72.3	76.0	**77.1**

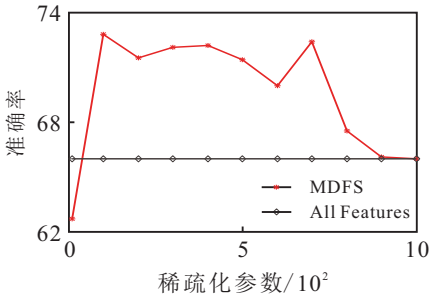
由表 2 和表 3 可以看出, 在 3 个数据库上, 本文算法的最高识别率 AC 和最高归一化信息 NMI 也超过其他几种算法. 另外, 在 3 个数据库上可以看出, 采用了类别标签信息的 TRC 算法和采用了稀疏局部保持的 MCFS 算法整体性能优于 VAR 和 LS 算法.

3.3　参数选择

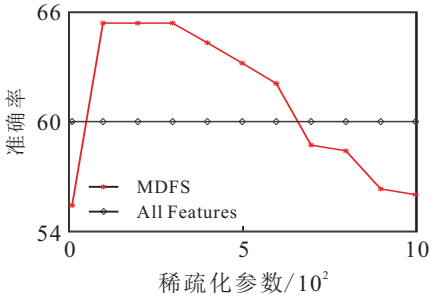
MDFS 模型中需要确定结构化稀疏参数 ε, 为了获得最优 ε 参数, 并且考察其对聚类结果的影响, 本文通过区间搜索方式在 ORL、COIL20、Isolet1 库上进行参数实验评估. 图 5 表示, 随着 ε 增加, MDFS 算法在各数据库上最优 AC 的变化曲线, 同时以 All Features 作为基准, 验证 ε 增加时 MDFS 算法的降维有效性. 从实验结果可以看出: 当 ε 取值较小时, 对应的 AC 较小, 无法达到 All Features 的实验结果; 当 ε 取值在 100～600 之间时, 在 3 个数据库上的实验结果较为理想; 而当 ε 取值较大时, AC 呈现下降趋势.



(a) ORL



(b) COIL20



(c) Isolet1

图 5　稀疏化参数评估

3.4 结果分析

由3.2节和3.3节的实验结果可以得到如下结论:

1) 采用聚类算法进行性能评估时, 本文提出的MDFS算法优于其他几种特征选择算法. 随着特征数的增加, 各算法聚类性能并不总是增加, 说明特征之间存在冗余. 在前100个特征内, 本文算法可以达到与采用所有特征(All Features)相似或更超越的实验结果, 验证了算法降维的有效性.

2) LS算法虽然考虑了样本几何流形结构, 但是由于其选择特征的方式为启发式逐个选取的策略, 不能保证特征子集最优, 因此其算法有效性不高. MCFS和TRC算法在3个数据库上的聚类性能较高, 原因在于MCFS算法在保持样本局部流形结构的基础上加入了L_1范数约束, 采用特征融合方式选取特征子集; TRC算法利用了样本标签信息进行特征子集选择. 而本文提出的MDFS算法融合了两种算法的优点, 通过样本标签构造类内和类间流形, 同时采用结构化稀疏约束降低了样本特征冗余, 提高了特征鉴别能力.

3) 当稀疏化参数ε较小时, 在目标函数中结构化稀疏项所占比重较少, 无法有效降低特征间冗余信息, 聚类效果较差; 而当ε过大时, 流形散度差所占比重降低, 也即样本标签信息无法得到有效利用, 同样会使得聚类性能降低. 本文通过区间搜索方式选择稀疏正则项参数, 使得标签信息和特征冗余信息达到最优平衡.

4 结 论

本文提出的MDFS算法通过标签信息刻画数据类内和类间流形结构, 以流形散度差为准则构造目标函数, 增强了特征鉴别能力, 同时对特征权重矩阵加以$L_{2,1}$范数约束, 形成特征竞争, 以减少特征冗余信息. 文中给出了MDFS模型的迭代求解算法、收敛性证明以及参数选择分析, 并在ORL库、COIL20库和Isolet1库上进行实验. 从实验结果可知, MDFS算法的准确率和归一化互信息明显优于VAR、LS、TRC、MCFS算法, 说明了MDFS算法的有效性. 但是, MDFS模型中稀疏化参数ε需要通过区间搜索得到最优值, 因此如何自适应地选择参数是今后研究的重点方向之一.

参考文献(References)

[1] Wu Xin-dong, Kui Yu, Wei Ding, et al. Online feature selection with streaming features[J]. Pattern Analysis and Machine Intelligence, 2012, 35(5): 1178-1192.

[2] Felipe A A, JoséL R, Alfredo R M, et al. Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection[J]. Expert Systems with Applications, 2012, 39(2): 1956-1967.

[3] Yang Yi, Ma Zhi-gang, Hauptmann A G, et al. Feature selection for multimedia analysis by sharing information among multiple tasks[J]. IEEE Trans on Multimedia, 2013, 15(3): 661-669.

[4] Liu J. Feature dimensionality reduction for myoelectric pattern recognition: A comparison study of feature selection and feature projection methods[J]. Medical Engineering and Physics, 2014, 36(12): 1716-1720.

[5] Fan B J, Cong Y, Du Y K. Discriminative multi-task objects tracking with active feature selection and drift correction[J]. Pattern Recognition, 2014, 47(12): 3828-3840.

[6] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-167.
(Yao X, Wang X D, Zhang Y X, et al. Summary of feature selection algorithms[J]. Control and Decision, 2012, 27(2): 161-167.)

[7] Guyon I, Jason W, Stephen B, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1/2/3): 389-422.

[8] Michalak K, Kwasnicka H. Correlation-based feature selection strategy in classification problems[J]. Int J of Applied Mathematics and Computer Science, 2006, 16(4): 503-511.

[9] He X F, Cai D, Niyogi P. Laplacian score for feature selection[C]. Advances in Neural Information Processing Systems. Vancouver, 2005: 507-514.

[10] Nie F P, Xiang S M, Jia Y Q, et al. Trace ratio criterion for feature selection[C]. Proc of the 23th AAAI Conf on Artificial Intelligence. Chicago: 2008: 671-676.

[11] Cai D, Zhang C Y, He X F. Unsupervised feature selection for multi-cluster data[C]. Advances in Neural Information Processing Systems. Washington, 2010: 333-342.

[12] Nie F P, Huang H, Cai X, et al. Efficient and robust feature selection via joint $L_{2,1}$-norms minimization[C]. Advances in Neural Information Processing Systems. Vancouver, 2010: 1812-1821.

[13] Hou C P, Nie F P, Li X L, et al. Joint embedding learning and sparse regression: A framework for unsupervised feature selection[J]. IEEE Trans on Cybernetics, 2014, 44(6): 793-804.

(责任编辑: 齐 霁)