

doi: 10.12052/gdutxb.210053

# 联合图嵌入与特征加权的无监督特征选择

张 巍, 张圳彬

(广东工业大学 计算机学院, 广东 广州 510006)

**摘要:** 在特征选择领域, 现有的大多数方法不能同时捕获不同特征有差异的权重, 不能对投影子空间施加正交约束来提高特征的判别力。为此, 本文提出联合图嵌入与特征加权的无监督特征选择方法(Joint Graph Embedding and Feature Weighting, JGEFW)。首先, 通过图嵌入局部结构学习获得相似度矩阵和聚类指示矩阵; 然后利用正交回归获得表征不同特征重要程度的权重矩阵, 以此选择出判别力强且非冗余的特征。此外, 本文还提出了一个交替迭代优化算法来求解JGEFW模型; 最后, 在4个数据集上进行实验验证。实验结果表明, JGEFW的聚类指标在大多数情况下优于其他对比算法。

**关键词:** 特征选择; 特征权重; 无监督学习; 图嵌入

中图分类号: TP391      文献标志码: A      文章编号: 1007-7162(2021)05-0016-08

## Joint Graph Embedding and Feature Weighting for Unsupervised Feature Selection

Zhang Wei, Zhang Zhen-bin

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

**Abstract:** In the area of feature selection, most existing methods cannot simultaneously capture the importance of different features and enforce orthogonal constraint on the projected subspace to improve the discrimination of selected features. To deal with this issue, a new unsupervised feature selection method called joint graph embedding and feature weighting (JGEFW) is proposed. To be specific, graph-embedding local structure learning is leveraged to obtain the similarity matrix and the cluster indicator matrix. Moreover, the weight matrix is learned through the orthogonal regression, to select the discriminative and non-redundant features. At last, an alternating iteration algorithm is developed for solving the proposed objective function. Finally, experimental results about clustering evaluation metrics on four publicly available datasets demonstrate that JGEFW is better than others in most cases.

**Key words:** feature selection; feature weight; unsupervised learning; graph embedding

伴随着信息科技的飞速发展, 在许多热点研究领域, 如图像、声音、文本、视频和基因工程等, 人们通常需要面对体量庞大的高维度数据<sup>[1-2]</sup>。为此, 降维方法中的特征选择(Feature Selection)是处理此类数据的常规方法。根据某些指标, 在不改变数据原始表示的前提下, 特征选择从高维数据中挑选信息量最为丰富的维度构成特征子集, 该过程各特征原始物理意义将得到维持<sup>[3]</sup>, 因此降维结果具备很好的解释性。在大部分情形下对全部特征进行采集将会是极

其耗时耗力高开销的或者是不可能的<sup>[4-5]</sup>, 特征选择只针对选择的特征进行采集, 故特征选择能减少待处理的数据量, 有助于进一步分析处理数据。

依据标签信息是否可用, 特征选择方法可分类为: 无监督(Unsupervised)、半监督(Semi-supervised)和有监督(Supervised)方法<sup>[6]</sup>。因为现实大部分数据缺乏标签, 对数据添加标签费力费时<sup>[7]</sup>。标签信息的缺乏使无监督学习难于其他两者。无监督特征选择可进一步分为3类: 过滤式(Filter)、包裹式(Wrapper)和嵌

入式(Embedding)<sup>[8]</sup>。此处主要介绍嵌入式。嵌入式方法将数据分布先验等数据属性纳入考虑,因此大多数情况下可以获得较优的特征信息。由于无监督特征选择具备捕获数据隐藏结构如局部流形结构和全局结构等的能力<sup>[9-10]</sup>,因此捕捉学习前者结构的嵌入式无监督特征选择通常能得到良好的效果<sup>[11]</sup>。

目前特征选择方法大多数采用嵌入式无监督的方式构建模型,但其中有些方法使得降维后的低维子空间特征判别性不强且存在冗余,导致选择的特征不具有较强的代表性,影响模型的性能。本文通过将自适应图嵌入学习获得的聚类指示矩阵在回归中进行拟合,并对低维的投影子空间施加正交约束,以此选择出判别性强且非冗余的特征,并通过一个表征特征重要程度的特征权重对角矩阵,进行特征选择并形成特征子集,进行后续任务处理。

## 1 国内外研究现状

国内外学者对于无监督特征选择方法进行了各种各样的探索。Cai等<sup>[12]</sup>提出多簇特征选择(Multi-Cluster Feature Selection, MCFS),使用 $l_1$ 范数正则化来维持数据的自然聚类的多簇结构;Qian等<sup>[13]</sup>提出鲁棒无监督特征选择(Robust Unsupervised Feature Selection, RUFS),在无监督聚类标签学习和特征学习过程同时利用最小化 $l_{2,1}$ 范数达到特征选择的目的;Hou等<sup>[3]</sup>提出了联合嵌入学习稀疏回归(Joint Embedding Learning and Sparse Regression, JELSR),使用谱聚类进行数据聚类结构学习,并使用系数回归在同一过程完成特征选择;Li等<sup>[14]</sup>提出了非负判别特征选择(Nonnegative Discriminative Feature Selection, NDFS),通过联合非负谱分析和稀疏回归统一学习以得到预想特征子集。

相较于以往方法,上述方法效果得到提升,但在降维过程中没有考虑对学习的投影子空间施加正交约束,以减少冗余特征的影响,提高选择特征的判别性。因此本文通过图嵌入局部结构学习,对数据间的相似图进行自适应调节,并将学习到的聚类指示矩阵作正交回归拟合,最后通过权重矩阵选择出判别性特征,该过程统一局部结构学习和特征选择在一个无监督框架中同时进行。

## 2 相关技术

此处介绍本文使用的符号及运算, $\mathbf{R}$ 表示实数集, $\mathbf{R}^{h_1 \times h_2}$ 表示共 $h_1$ 行且 $h_2$ 列的实数矩阵;以矩阵 $\mathbf{Q}, \mathbf{P} \in \mathbf{R}^{h_1 \times h_2}$ 为例, $\mathbf{Q}$ 的第 $i'$ 行第 $j'$ 列的元素表示为 $\mathbf{Q}_{i'j'}$ ,

$\mathbf{Q}$ 的迹表示为 $\text{Tr}(\mathbf{Q})$ , $\mathbf{Q}$ 的转置表示为 $\mathbf{Q}^T$ , $\mathbf{Q}$ 的F范数(Frobenius norm)表示为 $\|\mathbf{Q}\|_F = \sqrt{\sum_{j'=1}^{h_2} \sum_{i'=1}^{h_1} \mathbf{Q}_{i'j'}^2}$ ,其常见变换为 $\|\mathbf{Q}\|_F^2 = \text{Tr}(\mathbf{Q}^T \mathbf{Q}) = \text{Tr}(\mathbf{Q} \mathbf{Q}^T)$ , $\mathbf{Q}, \mathbf{P}$ 的哈达玛积(Hadamard product)为 $(\mathbf{Q} \circ \mathbf{P}) \in \mathbf{R}^{h_1 \times h_2}$ ,其第 $i'$ 行第 $j'$ 列的元素为 $(\mathbf{Q} \circ \mathbf{P})_{i'j'} = \mathbf{Q}_{i'j'} \mathbf{P}_{i'j'}$ 。

### 2.1 图嵌入局部结构学习

根据局部距离,图嵌入局部结构学习为各个数据点自适应分配最优近邻,以此来学习相似度矩阵,并挖掘潜在几何结构。

本文采用欧氏距离作为距离度量,默认向量为列向量。已知数据集 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbf{R}^{n \times d}$ ,其中 $n, d$ 分别表示样本数和特征维度; $\mathbf{S} \in \mathbf{R}^{n \times n}$ 表示相似度矩阵,其元素 $s_{ij}$ 表示 $\mathbf{x}_i, \mathbf{x}_j$ 之间的相似度, $\mathbf{s}_i \in \mathbf{R}^n$ 表示 $\mathbf{x}_i$ 与全体样本的相似度; $\mathbf{D}_S$ 表示关于 $\mathbf{S}$ 的度矩阵(Degree Matrix),其第 $i$ 个对角元素为 $\sum_j (s_{ij} + s_{ji})/2$ , $\mathbf{L}_S = \mathbf{D}_S - (\mathbf{S} + \mathbf{S}^T)/2$ 表示关于 $\mathbf{S}$ 的拉普拉斯矩阵(Laplacian Matrix)。

通过添加关于 $s_{ij}$ 的正则化项,使得 $n$ 个点都有 $1/n$ 的概率成为 $\mathbf{x}_i$ 的近邻,具体做法为对 $\mathbf{s}_i$ 施加近邻分配的先验约束 $\mathbf{s}_i^T \mathbf{1}_n = 1$ ,同时避免平凡解,其中 $\mathbf{1}_n$ 表示元素全为1的列向量。Nie等<sup>[15]</sup>通过对 $\mathbf{L}_S$ 施加秩约束 $\text{rank}(\mathbf{L}_S) = n - c$ ,限制学习的相似图连通分量为 $c$ ,该数值同时与自然聚类的簇个数完全相等,进而学习到更准确的数据潜在几何结构,如式(1)所示。

$$\begin{cases} \min_{\mathbf{S}} \sum_{i,j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \alpha s_{ij}^2 \\ \text{s.t. } \mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1, \text{rank}(\mathbf{L}_S) = n - c \end{cases} \quad (1)$$

由于 $\mathbf{S}$ 与 $\mathbf{L}_S, \mathbf{D}_S$ 直接相关,式(1)中的秩约束难以求解。Nie等<sup>[15]</sup>根据 $\text{rank}(\mathbf{L}_S) = n - c$ ,等价于理想情况下 $\mathbf{L}_S$ 具有前 $c$ 个和为0的特征值,记 $\sigma_v(\mathbf{L}_S)$ 表示 $\mathbf{L}_S$ 前 $v$ 个特征值,并由樊畿定理(Ky Fan's Theorem),得式(2)。

$$\sum_{v=1}^c \sigma_v(\mathbf{L}_S) = \min_{\mathbf{F} \in \mathbf{R}^{n \times c}, \mathbf{F}^T \mathbf{F} = \mathbf{I}_c} \text{Tr}(\mathbf{F}^T \mathbf{L}_S \mathbf{F}) \quad (2)$$

其中, $\mathbf{I}_c$ 为单位矩阵,规格为 $\mathbf{R}^{c \times c}$ , $\mathbf{F}$ 表示样本的低维嵌入,将式(2)迹约束替代式(1)秩约束,以便于求解,最终如式(3)所示,其中 $\alpha, \beta$ 均为正则化参数。

$$\begin{cases} \min_{\mathbf{S}, \mathbf{F}} \sum_{i,j=1}^n (\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \alpha s_{ij}^2) + 2\beta \text{Tr}(\mathbf{F}^T \mathbf{L}_S \mathbf{F}) \\ \text{s.t. } \mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1, \mathbf{F} \in \mathbf{R}^{n \times c}, \mathbf{F}^T \mathbf{F} = \mathbf{I}_c \end{cases} \quad (3)$$

### 2.2 特征加权正交回归

为了在投影子空间中保留关于数据样本更多的判别信息<sup>[16]</sup>并避免平凡解<sup>[17]</sup>,对最小二乘回归(Least

Square Regression, LSR)模型中的投影矩阵施加正交约束,得到正交回归模型(Orthogonal Regression)。对于该模型,  $\mathbf{W} \in \mathbf{R}^{d \times c}$  表示投影矩阵,  $\mathbf{b} \in \mathbf{R}^c$  表示偏置向量,  $\mathbf{Y} \in \mathbf{R}^{c \times n}$  表示数据样本标签矩阵,如式(4)所示。

$$\begin{cases} \min_{\mathbf{W}} \|\mathbf{X}\mathbf{W}\|^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y}\|_F^2 \\ \text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}_c \end{cases} \quad (4)$$

Wu等<sup>[18]</sup>在式(3)基础上添加表示特征权重矩阵  $\mathbf{\Omega} \in \mathbf{R}^{d \times d}$ , 且  $\mathbf{\Omega} = \text{diag}(\omega)$ ,  $\omega \in \mathbf{R}^d$  表示特征权重向量, 对其施加约束  $\omega^T \mathbf{1}_d = 1$ , 以消除权重取值大小带来的尺度影响。特征权重的大小表征各特征的重要性, 其物理意义为最小化正交回归中数据样本到拟合函数的垂直距离, 减少所选特征子集中的冗余信息, 如式(5)所示。最后选择  $\omega$  中权重排序位于前大  $\xi$  个的特征, 完成特征子集的生成以进行后续任务。

$$\begin{cases} \min_{\omega, \mathbf{W}} \|\mathbf{W}^T \mathbf{\Omega} \mathbf{X}^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y}\|_F^2 \\ \text{s.t. } \omega^T \mathbf{1}_d = 1, \omega \geq 0, \mathbf{W}^T\mathbf{W} = \mathbf{I}_c \end{cases} \quad (5)$$

由于正交回归可视为约束在施蒂费尔流形(Stiefel Manifold)  $\{\mathbf{W} \in \mathbf{R}^{d \times c} : \mathbf{W}^T\mathbf{W} = \mathbf{I}_c\}$  上的一个非均衡正交普洛克路斯忒斯问题(Unbalanced Orthogonal Procrustes Problem, UOPP), 其最优解很难得到<sup>[19]</sup>, 可借助Nie等提出的广义幂迭代算法<sup>[19]</sup>(Generalized Power Iteration, GPI)进行求解。

### 3 JGEFW模型

#### 3.1 模型构建

将式(3)与式(5)联立, 得到本文的目标函数, 参见式(6)。式(6)将图嵌入局部结构学习与特征加权正交回归结合起来, 获得一个统一的目标函数。其中: 第一、二部分为图嵌入局部结构学习, 自适应学习局部流形结构, 即数据样本间的相似度矩阵; 第三部分约束学习到的数据之结构图的连通分量个数与自然聚类的簇个数完全相等, 以学习到更好的数据样本图结构; 第四部分将局部流形结构学习得到的聚类指示矩阵在正交回归中拟合并获得权重矩阵。

进行联合学习是为了保留更多的判别信息并得到表征各个特征重要性的权重矩阵。整体框架模型参见式(6)。

$$\begin{cases} \min_{\mathbf{S}, \mathbf{F}, \mathbf{W}, \omega} \sum_{i,j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \alpha s_{ij}^2 + \\ 2\beta \text{Tr}(\mathbf{F}^T \mathbf{L}_S \mathbf{F}) + \gamma \|\mathbf{W}^T \mathbf{\Omega} \mathbf{X}^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{F}\|_F^2 \\ \text{s.t. } \mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1, \omega^T \mathbf{1}_d = 1, \omega \geq 0, \\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_c, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \end{cases} \quad (6)$$

#### 3.2 模型求解

根据变量  $\mathbf{b}$  的极值条件, 有  $\mathbf{b} = (1/n)(\mathbf{F}^T \mathbf{1}_n - \mathbf{W}^T \mathbf{\Omega} \mathbf{X}^T \mathbf{1}_n)$ , 其中  $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{1}_n \mathbf{1}_n^T \in \mathbf{R}^{n \times n}$  为中心化矩阵, 故目标函数可简化为

$$\begin{cases} \min_{\mathbf{S}, \mathbf{F}, \mathbf{W}, \omega} \sum_{i,j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \alpha s_{ij}^2 + \\ 2\beta \text{Tr}(\mathbf{F}^T \mathbf{L}_S \mathbf{F}) + \gamma \|\mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W} - \mathbf{H} \mathbf{F}\|_F^2 \\ \text{s.t. } \mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1, \omega^T \mathbf{1}_d = 1, \omega \geq 0, \\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_c, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \end{cases} \quad (7)$$

(1) 更新  $\mathbf{F}$ 。对  $\mathbf{F}$  进行优化求解, 变换得

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \text{Tr}(\mathbf{F}^T (2\beta \mathbf{L}_S + \gamma \mathbf{H}) \mathbf{F} - 2\mathbf{F}^T (\gamma \mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W}))$$

等价于求解

$$\begin{cases} \min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \text{Tr}(\mathbf{F}^T \mathbf{A}_1 \mathbf{F} - 2\mathbf{F}^T \mathbf{B}_1) \\ \mathbf{A}_1 = 2\beta \mathbf{L}_S + \gamma \mathbf{H}, \mathbf{B}_1 = \gamma \mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W} \end{cases} \quad (8)$$

此归属UOPP问题, 可由广义幂迭代算法<sup>[19]</sup>求解  $\mathbf{F}$ 。

(2) 更新  $\mathbf{S}$ 。根据文献<sup>[15]</sup>, 由于对于不同的样本其求解是独立的, 故以样本  $\mathbf{x}_i$  的相似度向量  $\mathbf{s}_i$  求解为例进行说明。固定其他变量, 得

$$\min_{\mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1} \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \alpha s_{ij}^2 + 2\beta \text{Tr}(\mathbf{F}^T \mathbf{L}_S \mathbf{F})$$

等价于求解

$$\min_{\mathbf{s}_i^T \mathbf{1}_n = 1, 0 \leq s_i \leq 1} \|\mathbf{s}_i + (1/2\alpha) \mathbf{u}_i\|_2^2 \quad (9)$$

其中,  $\mathbf{u}_i = [u_{i1}; \dots; u_{in}]$ ,  $u_{ij} = u_{ij}^x + \beta u_{ij}^f$ ,  $u_{ij}^x = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ ,  $u_{ij}^f = \|\mathbf{f}_i - \mathbf{f}_j\|_2^2$ 。该优化问题可归结为近端问题<sup>[20]</sup>(Proximal Problem)求解。

(3) 更新  $\mathbf{W}$ 。变换得

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_c} \text{Tr}(\mathbf{W}^T \mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W} - 2\mathbf{W}^T \mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{F})$$

等价于求解

$$\begin{cases} \min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_c} \text{Tr}(\mathbf{W}^T \mathbf{A}_2 \mathbf{W} - 2\mathbf{W}^T \mathbf{B}_2) \\ \mathbf{A}_2 = \mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{X} \mathbf{\Omega}, \mathbf{B}_2 = \mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{F} \end{cases} \quad (10)$$

同理, 求解式(11)属于UOPP问题, 可由广义幂迭代算法求得  $\mathbf{W}$ 。

(4) 更新  $\omega$ 。

$$\begin{aligned} \min_{\omega^T \mathbf{1}_d = 1, \omega \geq 0} \|\mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W} - \mathbf{H} \mathbf{F}\|_F^2 = \\ \text{Tr}(\mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{X} \mathbf{\Omega} \mathbf{W} \mathbf{W}^T) - \text{Tr}(2\mathbf{\Omega} \mathbf{X}^T \mathbf{H} \mathbf{F} \mathbf{W}^T) = \\ \omega^T ((\mathbf{X}^T \mathbf{H} \mathbf{X})^T \circ (\mathbf{W} \mathbf{W}^T)) \omega - \omega^T \text{diag}(2\mathbf{X}^T \mathbf{H} \mathbf{F} \mathbf{W}^T) \end{aligned}$$

原问题转换为

$$\begin{cases} \min_{\omega^T \mathbf{1}_d=1, \omega \geq 0} \omega^T \mathbf{Z} \omega - \omega^T \mathbf{z} \\ \mathbf{Z} = (\mathbf{X}^T \mathbf{H} \mathbf{X}) \circ (\mathbf{W} \mathbf{W}^T) \\ \mathbf{z} = \text{diag}(2\mathbf{X}^T \mathbf{H} \mathbf{F} \mathbf{W}^T) \end{cases} \quad (11)$$

其中,  $\text{diag}(\mathbf{Q})$ 表示以向量 $\mathbf{Q}$ 的元素为对角元素的方阵。由增广拉格朗日乘子法<sup>[18]</sup>(Augmented Lagrangian Multiplier Method, ALM),可求解得 $\omega$ 。

此处介绍更新 $\mathbf{F}, \mathbf{W}$ 所需的广义幂迭代算法,以求解类似  $\min_{\mathbf{Z}^T \mathbf{Z} = \mathbf{I}_{\epsilon \times \epsilon}} \text{Tr}(\mathbf{Z}^T \mathbf{A} \mathbf{Z} - 2\mathbf{Z}^T \mathbf{B})$  的优化问题,其中对称矩阵 $\mathbf{A} \in \mathbf{R}^{\epsilon \times \epsilon}$ 和矩阵 $\mathbf{B} \in \mathbf{R}^{\epsilon \times \epsilon}$ ,如式(8)和式(10)所示,并输出矩阵 $\mathbf{Z} \in \mathbf{R}^{\epsilon \times \epsilon}$ ,相关流程如下所示。

初始化:随机矩阵 $\mathbf{Z}$ ,满足 $\mathbf{Z}^T \mathbf{Z} = \mathbf{I}_{\epsilon \times \epsilon}$

参数 $\eta$ 使得 $\tilde{\mathbf{A}} = \eta \mathbf{I}_d - \mathbf{A} \in \mathbf{R}^{\epsilon \times \epsilon}$ 为正定矩阵

过程:

① 更新 $\mathbf{Z} \leftarrow 2\tilde{\mathbf{A}}\mathbf{Z} + 2\mathbf{B}$

② 对 $\mathbf{Z}$ 进行奇异值分解(Singular Value Decomposition, SVD),计算 $\mathbf{U}\mathbf{Z}\mathbf{V}^T = \mathbf{Z}, \mathbf{U} \in \mathbf{R}^{\epsilon \times \epsilon}, \mathbf{Z}, \mathbf{V} \in \mathbf{R}^{\epsilon \times \epsilon}$

③ 更新 $\mathbf{Z} \leftarrow \mathbf{U}\mathbf{V}^T$

重复步骤①~③直到收敛。

基于上述算法,交替更新 $\mathbf{F}, \mathbf{S}, \mathbf{W}, \omega$ ,直到收敛。

## 4 实验

为验证本文提出的JGEFW的有效性,根据文献[21]将JGEFW与几种常见的无监督特征选择算法,包括拉普拉斯分数(Laplacian Score, LS)、多簇特征选择(MCFS)、无监督判别特征选择(Unsupervised Discriminative Feature Selection, UDFS)、双自表示和流形正则化的鲁棒无监督特征选择(Dual Self-representation and Manifold Regularization, DSRMR)以及包含所有特征的基线模型(Baseline)进行对比。

根据文献[21]选择在4个公开数据集YALE、TOX、ISOLET和COIL上进行比较,且设定选择特征数范围均为{20, 30, 40, 50, 60, 70, 80, 90, 100}。对于不同的选择特征数均采用30次 $k$ 均值( $k$ -means)聚类取平均值以减少随机效应的影响,并记录各评估指标的最优结果来表征对所选择的特征的评估。实验设备具体配置为Core(TM) i9-9900K CPU@3.60GHz, RAM 32GB。

### 4.1 评估指标

实验评估指标包括聚类准确率(Accuracy, ACC)和标准化互信息(Normalized Mutual Information, NMI)。这两者取值范围均为[0, 1],且数值越大代表聚类效果越好。

$$\text{ACC} = \frac{1}{n} \sum_i \delta(r_i, \text{map}(g_i)) \quad (12)$$

其中, $n$ 为数据样本数, $r_i, g_i$ 分别是关于 $\mathbf{x}_i$ 的真实标签和预测聚类标签; $\delta(\tau, \varphi)$ 为指示函数(Indicator),若 $\tau = \varphi$ 则 $\delta(\tau, \varphi) = 1$ ,否则 $\delta(\tau, \varphi) = 0$ ;使用KM算法(Kuhn-Munkres)<sup>[22]</sup>的最优映射函数 $\text{map}()$ 将聚类标签映射到相应真实标签。

$$\text{MI} = \sum_{\lambda_{\mu_1} \in \mathbf{A}, \lambda'_{\mu_2} \in \mathbf{A}'} p(\lambda_{\mu_1}, \lambda'_{\mu_2}) \log \frac{p(\lambda_{\mu_1}, \lambda'_{\mu_2})}{p(\lambda_{\mu_1})p(\lambda'_{\mu_2})} \quad (13)$$

其中, $\mathbf{A}, \mathbf{A}'$ 分别为真实标签集和聚类标签集,MI为两者之间的互信息(Mutual Information), $p(\lambda_{\mu_1}), p(\lambda'_{\mu_2})$ 分别表示样本分属于簇 $\lambda_{\mu_1}, \lambda'_{\mu_2}$ 的概率, $p(\lambda_{\mu_1}, \lambda'_{\mu_2})$ 表示样本同属于簇 $\lambda_{\mu_1}, \lambda'_{\mu_2}$ 的概率。

$$\text{NMI}(\mathbf{A}, \mathbf{A}') = \frac{\text{MI}}{\sqrt{E(\mathbf{A})E(\mathbf{A}')}} \quad (14)$$

其中, $E(\mathbf{A}), E(\mathbf{A}')$ 分别表示 $\mathbf{A}, \mathbf{A}'$ 的熵(Entropy),NMI反映真实标签与聚类标签之间的一致性关系。

### 4.2 数据集介绍

实验采用的4个公开数据集简介,如表1所示。

表1 数据集简要信息

Table 1 Brief information of datasets

| 数据集    | 样本数   | 特征维度  | 类别数 | 类型   |
|--------|-------|-------|-----|------|
| YALE   | 165   | 1 024 | 15  | 人脸图像 |
| TOX    | 171   | 5 748 | 4   | 微阵列  |
| ISOLET | 1 560 | 617   | 26  | 混合特征 |
| COIL   | 1 440 | 1 024 | 20  | 目标图像 |

YALE数据集属于人脸图像数据集,其包含15个人共165张的灰度图像,每个人包含11张不同面部表情和环境的图像,包括中心光、左光、右光、快乐、正常、悲伤、困倦、惊讶、戴眼镜和眨眼<sup>[23]</sup>。

TOX数据集属于生物数据集,共包含4个类别的171个样本,每个样本维度为5 748维。

ISOLET数据集是从30个测试者每人读出两次全部英文字母收集而来,样本共1 560个,维度为617。

COIL数据集包含1 440张图像,共20类物品,每个物品有72张图像,其来源于相机以固定间隔5°的角度拍摄得到每类物品的图像,裁剪大小为32×32像素,故维度为1 024维。

### 4.3 各算法参数设置

根据文献[21],设置所有对比算法相似度矩阵邻关系由 $k$ 近邻( $k$ -Nearest Neighbors, kNN)生成且 $k$ 设

置为5;LS近邻权重由余弦相似度生成;MCFS近邻权重由Binary生成且nUseEigenfunction设置为1;UDFS和DSRMR的正则化参数设置为1;对于本算法JGEFW,设置 $k$ 为15,步长参数为1.1,惩罚参数(Penalty Parameter)为1 000,根据文献[24]对 $F$ 进行初始化,同时根据文献[15]对 $\alpha$ 进行初始化,并对参数 $\beta,\gamma$ 进行网格搜索(Grid Search),搜索范围均为 $\{10^{-3}, 10^{-1}, 10^1, 10^3, 10^5\}$ 。

4.4 实验结果与分析

4.4.1 评估指标最优结果

表2~3分别展示几种无监督特征选择算法在各数据集上,经参数搜索后选取最优参数且选取各特征选择数情况下的最佳聚类结果及相应的标准差,同时表2~3最后一行为各个算法在全部数据集上的平均聚类效果。表中加粗项表示在该数据集效果最优,下划线项表示在该数据集上效果次优,对比算法结果均由文献[21]给出。

表 2 几种无监督特征选择算法在各数据集聚类准确率(%)±标准差(%)比较

| Table 2 Results on clustering evaluation metric of ACC(%)±std(%) of all compared unsupervised feature selection methods on all datasets |                    |            |            |                    |                    |                    |
|-----------------------------------------------------------------------------------------------------------------------------------------|--------------------|------------|------------|--------------------|--------------------|--------------------|
| 数据集                                                                                                                                     | LS                 | MCFS       | UDFS       | DSRMR              | Baseline           | Ours               |
| YALE                                                                                                                                    | <u>36.57</u> ±2.34 | 33.35±1.97 | 32.14±1.96 | 33.62±2.09         | 34.30±3.50         | <b>39.58</b> ±2.63 |
| TOX                                                                                                                                     | 43.04±2.66         | 41.48±2.48 | 42.09±2.53 | 42.32±2.60         | <u>44.21</u> ±3.48 | <b>44.58</b> ±2.32 |
| ISOLET                                                                                                                                  | 52.54±3.18         | 56.80±3.68 | 39.74±1.66 | 53.02±2.94         | <u>59.06</u> ±4.48 | <b>62.94</b> ±3.27 |
| COIL                                                                                                                                    | 52.17±3.50         | 53.48±4.20 | 47.22±2.68 | <u>55.39</u> ±3.48 | 52.98±3.99         | <b>57.25</b> ±3.52 |
| 平均效果                                                                                                                                    | 46.08              | 46.28      | 40.30      | 46.09              | <u>47.64</u>       | <b>51.09</b>       |

表 3 几种无监督特征选择算法在各数据集标准化互信息(%)±标准差(%)比较

| Table 3 Results on clustering evaluation metric of NMI(%)±std(%) of all compared unsupervised feature selection methods on all datasets |                    |                    |            |            |                    |                    |
|-----------------------------------------------------------------------------------------------------------------------------------------|--------------------|--------------------|------------|------------|--------------------|--------------------|
| 数据集                                                                                                                                     | LS                 | MCFS               | UDFS       | DSRMR      | Baseline           | Ours               |
| YALE                                                                                                                                    | <u>42.92</u> ±1.94 | 40.20±1.56         | 38.93±1.73 | 41.21±1.62 | 42.07±2.71         | <b>45.04</b> ±1.92 |
| TOX                                                                                                                                     | 12.87±2.33         | 11.75±2.09         | 11.98±0.61 | 12.66±1.42 | <b>16.73</b> ±5.54 | <u>15.42</u> ±3.51 |
| ISOLET                                                                                                                                  | 70.77±1.05         | 73.34±1.30         | 54.82±1.06 | 69.53±1.34 | <b>77.27</b> ±1.95 | <u>75.55</u> ±1.37 |
| COIL                                                                                                                                    | 67.39±1.29         | <u>71.12</u> ±1.61 | 61.51±1.50 | 70.62±1.90 | <b>72.04</b> ±1.47 | 70.24±2.47         |
| 平均效果                                                                                                                                    | 48.49              | 49.10              | 41.81      | 48.51      | <b>52.03</b>       | <u>51.56</u>       |

由表2可知,本方法在各数据集和平均聚类准确率上效果均为最优,同时在YALE、ISOLET、COIL和平均聚类准确率与次优效果相比均有较大的提升。由表3可知,本方法在YALE上效果为最优,且与次优效果相比有较大的提升,而在TOX、ISOLET和平均标准化互信息为次优,由于Baseline模型对全部特征进行利用而不加以选择,因此对于某些包含噪声和冗余特征较少的数据集而言,该做法有时也能获得较好的效果,但并无包含降维操作。

4.4.2 评估指标随各特征选择数的变化

图1~2分别表示在不同聚类指标下几种不同的特征选择算法在各数据集上选择不同特征数时的聚类效果。图中横坐标表示不同的选择选择数,纵坐标分别表示以聚类准确率(ACC,%)和标准化互信息(NMI,%)作为聚类指标时的值。由于文献[21]并无给出对比算法在各选择特征数的实际数值,故该部分数据是根据该论文设置,实际运行代码获得的,得到

的各对比算法实际最优效果允许与表2~3存在差距。

由图1可知,本方法在YALE上聚类准确率变动较为波动,与其他对比算法情况相似,且在大多数选择特征数范围优于其他对比算法;在TOX上聚类准确率大致呈整体上升趋势,且在大多数选择特征数范围效果均优于LS、MCFS和DSRMR;在ISOLET上随着选择特征数增加,效果大致呈上升趋势,与大多数对比算法变化趋势一致,且在选择特征数为90时聚类准确率为所有对比算法下的最优;在COIL上随着选择特征数增加,聚类准确率在多数选择特征数范围均优于其他对比算法,且在选择特征数为70时聚类准确率为所有对比算法下的最优。

由图2可知,本方法在YALE上标准化互信息变动较为波动,与其他对比算法情况相似,但在选择特征数为30时标准化互信息为所有对比算法下的最优,表明JGEFW在选择特征数较小时便能选择出判别性强的信息;在TOX上标准化互信息大致呈上升

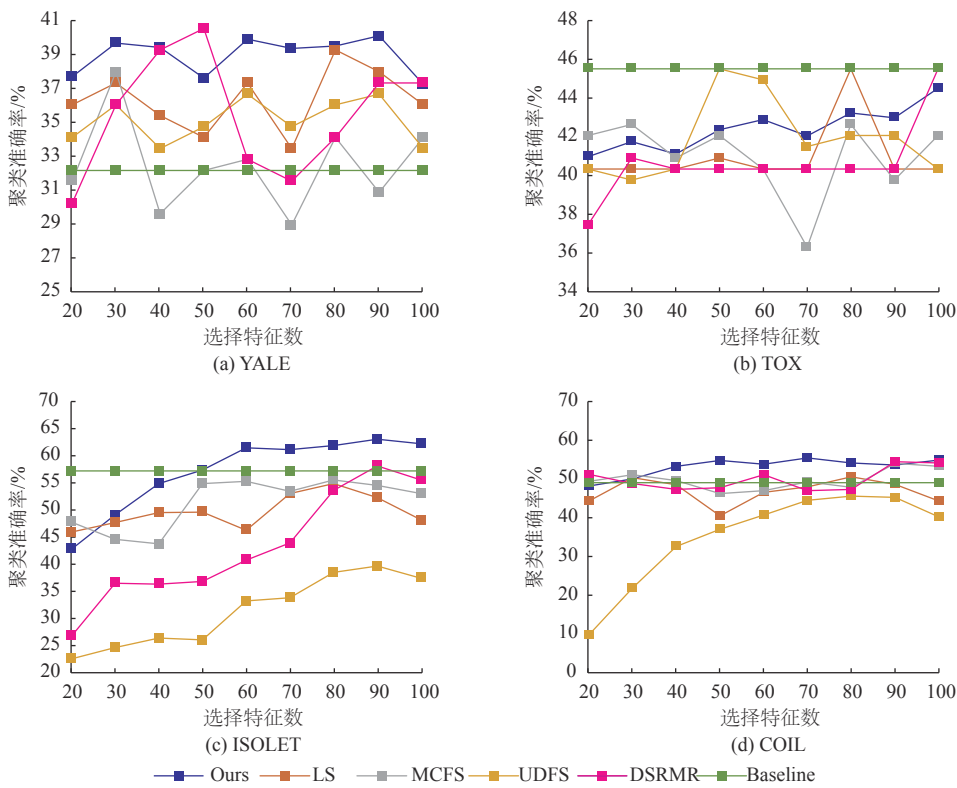


图 1 各数据集几种算法不同特征选择数的聚类准确率(%)比较

Fig.1 Results on clustering evaluation metric of ACC (%) of different numbers of selected features of all compared unsupervised feature selection methods on all datasets

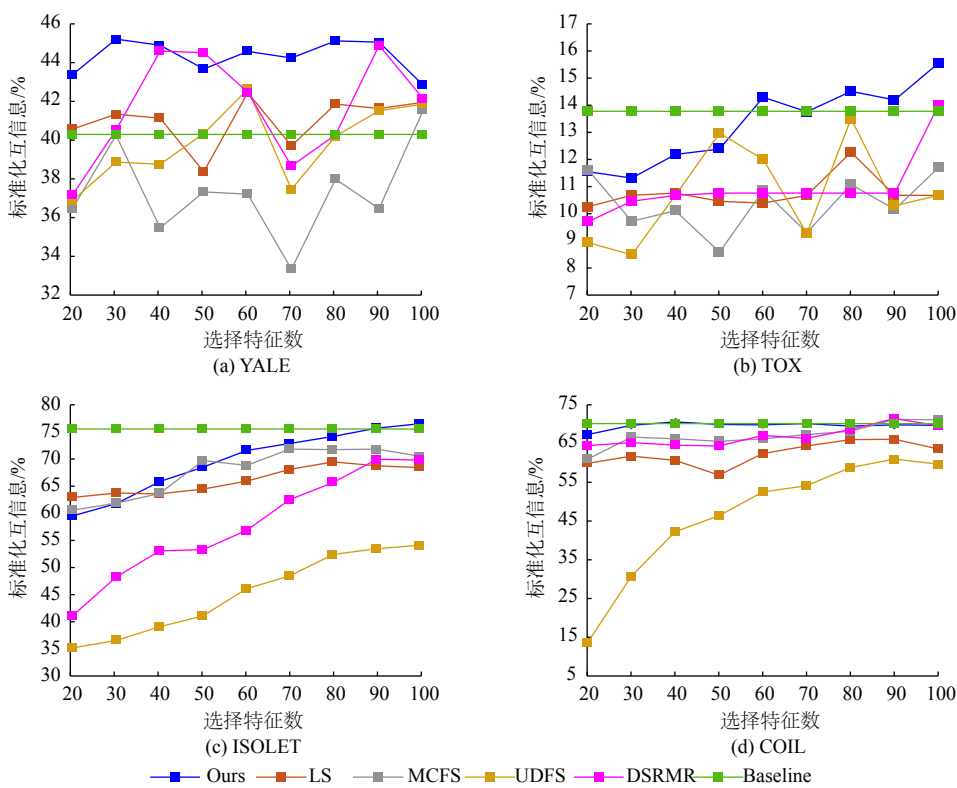


图 2 各数据集几种算法不同特征选择数的标准化互信息(%)比较

Fig.2 Results on clustering evaluation metric of NMI (%) of different numbers of selected features of all compared unsupervised feature selection methods on all datasets

趋势,与其他对比算法趋势完全不同,且当选择特征数在大于70的范围内,标准化互信息均为所有对比算法下的最优;在ISOLET上随着选择特征数增加,标准化互信息大致呈提升趋势,与其他绝大多数对比算法变化趋势相同,且在选择特征数为100时标准化互信息优于Baseline的效果;在COIL上随着选择特征数增加,标准化互信息在Baseline算法上下波动,表明本算法能有效学习到数据间的潜在几何结构,并对判别力强的特征加以选择。

4.4.3 参数敏感性分析

为探究算法的性能,进行参数敏感性分析实验

(Parameter Sensitivity),以考虑各参数变动对模型的影响。此处以COIL数据集为例,在最优聚类准确率效果中先固定 $\beta = 10$ ,变化 $\gamma$ ,观察算法的聚类准确率变化,再固定 $\gamma = 0.1$ ,变化 $\beta$ ,观察算法的聚类准确率变化;在最优标准化互信息效果中先固定 $\beta = 10^3$ ,变化 $\gamma$ ,观察算法的标准化互信息变化,再固定 $\gamma = 10^5$ ,变化 $\beta$ ,观察算法的标准化互信息变化。

由图3可知,JGEFW的聚类准确率随着选择特征数、 $\beta$ 和 $\gamma$ 的改变略有波动,而JGEFW的标准化互信息则随着选择特征数、 $\beta$ 和 $\gamma$ 的改变总体上较为平稳,具有一定的鲁棒性。

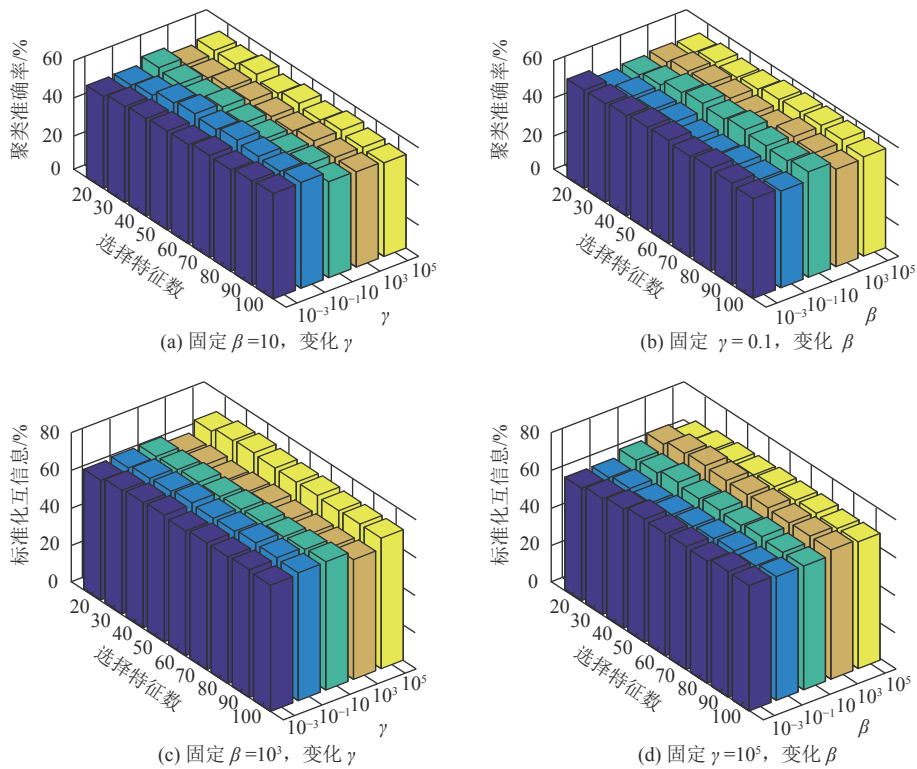


图3 本方法在COIL的参数敏感性分析实验结果

Fig.3 Parameters sensitivity analysis about clustering evaluation metric of the proposed method on COIL dataset

4.4.4 收敛性实验证明

此处以COIL数据集为例,其他数据集上的收敛情况与此相似,目标函数值的收敛曲线如图4所示。由图4可知,随着迭代次数逐渐增加,目标函数值不断减少,最后趋于一个稳定值,表明该算法是收敛的。

5 结束语

本文给出了一种改进的无监督特征选择算法,称为JGEFW。通过局部结构学习自适应地调节数据的相似图,使其满足自然聚类的簇的要求,提高图构造的准确度。并通过加权特征回归的方法,使图构造

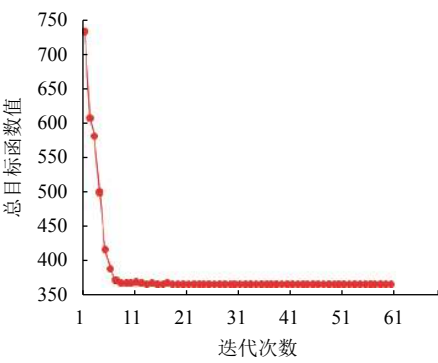


图4 本方法在COIL数据集上的收敛结果示意

Fig.4 Result of convergence experiments of the proposed method on COIL dataset

的聚类指示矩阵在正交回归中拟合以学习权重矩阵,并挑选出权重值大的判别特征和非冗余特征形成特征子集。实验结果表明,JGEFW的性能在平均情况下优于其他对比的特征选择方法。未来工作将尝试提高模型的聚类性能和迭代速度。

## 参考文献:

- [1] HU J, LI Y, GAO W, *et al.* Robust multi-label feature selection with dual-graph regularization [J]. [Knowledge-Based Systems](#), 2020, 203: 106126.
- [2] 费伦科, 秦建阳, 滕少华, 等. 近似最近邻大数据检索哈希散列方法综述[J]. [广东工业大学学报](#), 2020, 37(3): 23-35. FEI L K, QIN J Y, TENG S H, *et al.* Hashing for approximate nearest neighbor search on big data: A survey [J]. [Journal of Guangdong University of Technology](#), 2020, 37(3): 23-35.
- [3] HOU C, NIE F, YI D, *et al.* Feature selection via joint embedding learning and sparse regression[C]//IJCAI International Joint Conference on Artificial Intelligence. Spain: Morgan Kaufmann, 2011: 1324-1329.
- [4] WANG S, WANG H. Unsupervised feature selection via low-rank approximation and structure learning [J]. [Knowledge-Based Systems](#), 2017, 124: 70-79.
- [5] 刘艳芳, 李文斌, 高阳. 基于自适应邻域嵌入的无监督特征选择算法[J]. [计算机研究与发展](#), 2020, 57(8): 1639-1649. LIU Y F, LI W B, GAO Y. Adaptive neighborhood embedding based unsupervised feature selection [J]. [Journal of Computer Research and Development](#), 2020, 57(8): 1639-1649.
- [6] 滕少华, 冯镇业, 滕璐瑶, 等. 联合低秩表示与图嵌入的无监督特征选择[J]. [广东工业大学学报](#), 2019, 36(5): 7-13. TENG S H, FENG Z Y, TENG L Y, *et al.* Joint low-rank representation and graph embedding for unsupervised feature selection [J]. [Journal of Guangdong University of Technology](#), 2019, 36(5): 7-13.
- [7] ZHENG W, YAN H, YANG J. Robust unsupervised feature selection by nonnegative sparse subspace learning [J]. [Neurocomputing](#), 2019, 334: 156-171.
- [8] DING D, YANG X, XIA F, *et al.* Unsupervised feature selection via adaptive hypergraph regularized latent representation learning [J]. [Neurocomputing](#), 2020, 378: 79-97.
- [9] TENG L, FENG Z, FANG X, *et al.* Unsupervised feature selection with adaptive residual preserving [J]. [Neurocomputing](#), 2019, 367: 259-272.
- [10] LIU X, WANG L, ZHANG J, *et al.* Global and local structure preservation for feature selection [J]. [IEEE Transactions on Neural Networks and Learning Systems](#), IEEE, 2014, 25(6): 1083-1095.
- [11] DU L, SHEN Y D. Unsupervised feature selection with adaptive structure learning[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015-August: 209-218.
- [12] CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 333-342.
- [13] QIAN M, ZHAI C. Robust unsupervised feature selection[C]//IJCAI International Joint Conference on Artificial Intelligence. Beijing: Morgan Kaufmann, 2013: 1621-1627.
- [14] LI Z, YANG Y, LIU J, *et al.* Unsupervised feature selection using nonnegative spectral analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Toronto: AAAI, 2012: 1026-1032.
- [15] NIE F, WANG X, HUANG H. Clustering and projected clustering with adaptive neighbors[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 977-986.
- [16] XU X, WU X, WEI F, *et al.* A general framework for feature selection under orthogonal regression with global redundancy minimization [J]. [IEEE Transactions on Knowledge and Data Engineering](#), 2021, 99: 1-1.
- [17] ZHAO H, WANG Z, NIE F. Orthogonal least squares regression for feature extraction [J]. [Neurocomputing](#), 2016, 216: 200-207.
- [18] WU X, XU X, LIU J, *et al.* Supervised feature selection with orthogonal regression and feature weighting [J]. [IEEE Transactions on Neural Networks and Learning Systems](#), 2020, 32(5): 1831-1838.
- [19] NIE F, ZHANG R, LI X. A generalized power iteration method for solving quadratic problem on the Stiefel manifold [J]. [Science China\(Information Sciences\)](#), 2017, 60(11): 5-11.
- [20] HUANG J, NIE F, HUANG H. A new simplex sparse learning model to measure data similarity for clustering[C]//IJCAI International Joint Conference on Artificial Intelligence. Buenos Aires: Morgan Kaufmann, 2015: 3569-3575.
- [21] LIU Y, YE D, LI W, *et al.* Robust neighborhood embedding for unsupervised feature selection [J]. [Knowledge-Based Systems](#), 2020, 193: 105462.
- [22] LI X, ZHANG H, ZHANG R, *et al.* Generalized uncorrelated regression with adaptive graph for unsupervised feature selection [J]. [IEEE Transactions on Neural Networks and Learning Systems](#), 2019, 30(5): 1587-1595.
- [23] THARWAT A, GABER T, IBRAHIM A, *et al.* Linear discriminant analysis: a detailed tutorial [J]. [AI Communications](#), 2017, 30(2): 169-190.
- [24] SHANG R, WANG W, STOLKIN R, *et al.* Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection [J]. [IEEE Transactions on Cybernetics](#), IEEE, 2018, 48(2): 793-806.

(责任编辑:王威娜)