doi:10.12068/j.issn.1005-3026.2024.12.005

基于稀疏自表示及流形正则化的无监督特征选择

刘 杰,谭文静,李占山
(吉林大学 计算机科学与技术学院,吉林 长春 130012;)

摘 要:基于自表示的无监督特征选择能够处理未标记数据且不受伪标签影响.为了令此类方法同时具有良好的鲁棒性、保留样本局部结构、能选出最具代表性的特征,提出了一种新的方法,并设计了一个对应的迭代优化算法来计算其目标函数.该方法先对样本异常值进行识别和处理,然后将传统的自表示模型与非凸稀疏约束和流形正则结合形成目标模型,再将预处理后的数据放入模型进行特征选择,最后使用所选特征进行聚类.将所提方法在9个真实数据集上与7种方法进行对比实验,实验结果表明,所提方法可以有效解决无监督特征选择问题.

关 键 词:无监督特征选择;自表示;鲁棒;稀疏;流形正则化
中图分类号:TP 181 **文献标志码**:A **文章编号**:1005-3026(2024)12-1706-11

Unsupervised Feature Selection Based on Sparse Self-representation with Manifold Regularization

LIU Jie, *TAN Wen-jing*, *LI Zhan-shan*
(College of Computer Science and Technology, Jilin University, Changchun 130012, China. Corresponding author: LI Zhan-shan, E-mail: lizs@jlu.edu.cn)

Abstract: Self-representation based unsupervised feature selection can handle unlabeled data without being affected by pseudo-labeling. To ensure that such methods simultaneously achieve good robustness, preserve the local structure of samples, and select the most representative features, a new approach is proposed, and a corresponding iterative optimization algorithm is designed to compute its objective function. The method first identifies and processes outliers of samples, then combines the traditional self-representation model with non-convex sparse constraint and manifold regularization to form the target model, and puts the preprocessed data into the model for feature selection. Finally, the method uses the selected features for clustering. The proposed method is compared with seven methods on nine real data sets for experiments, and the experimental results show that the proposed method can effectively solve the unsupervised feature selection problem.

Key words: unsupervised feature selection; self-representation; robust; sparse; manifold regularization

在机器学习、数据挖掘等领域,特征选择有助于减少维数诅咒的影响,提高预测性能.根据数据是否被标记,特征选择方法分为有监督的、半监督的和无监督的[1].数据标签在现实中通常不可用或成本较高,因此没有先验知识也表现较好的无监督特征选择(unsupervised feature selection,UFS)得到了广泛关注.

就选择策略而言,特征选择方法可分成3种类型:过滤式[2-5]、包裹式[6]和嵌入式[7-9].嵌入式算法结合了前两种方法的优点,在一定程度上统一了特征选择和学习算法,可以获得更好的结果.GSR_SFS(graph self-representation sparse feature selection)[7]将子空间学习与稀疏的特征级自表示(self-representation,SR)方法相结合,使特征选择

收稿日期:2023-06-09
基金项目:国家自然科学基金资助项目(62276060);吉林省发展和改革委员会资助项目(2019C053-9).
作者简介:刘 杰(1973-),女,吉林长春人,吉林大学副教授;李占山(1966-),男,吉林长春人,吉林大学教授,博士生导师.

过程更稳定并获得更好的解释能力.SOCFS(structured optimal graph feature selection)[8]在进行特征选择和局部结构学习时自适应地确定相似性矩阵,并使用$l_{2,1}$范数来实现稀疏性.在UFS领域的嵌入式方法中,有专注于数据全局结构的[9],但大多数方法都会关注数据的局部几何结构,该结构在无监督情况下已被证明比全局结构更关键[10].

根据使用的基本模型,UFS的嵌入式方法可以分为两类[11].一类是基于回归的模型,其本质上是通过学习样本的伪标签将UFS转变为有监督的特征选择,如SCFS(subspace clustering unsupervised feature selection)[12]在一个集成的框架中应用子空间学习、聚类分析和稀疏学习,采用矩阵分解得到聚类标签矩阵,并利用回归模型优化系数矩阵选择最重要的特征.另一类是基于表示学习的模型,其核心思想是学习数据的低维表征,为特定的任务(如分类)保留最重要的信息,可以避免不准确的伪标签导致选择的特征不可靠[11].

作为表示学习的一个特例,基于SR的UFS方法利用原始数据固有的自我表示特性,对高维无标签数据进行降维.RSR(regularized self-representation)[13]是这类方法中的代表之一,它为了选择更具代表性的特征并提高算法的鲁棒性,对误差函数和稀疏项同时使用$l_{2,1}$范数.近些年基于RSR提出了各种改进算法.NOVRSR(non-convex regularized self-representation)[14]相较于RSR使用了效果更好的$l_{2,1-2}$范数[15]作为稀疏项,并验证了其有效性,但该方法没有考虑到样本原有的结构信息.基于NOVRSR,本文提出了一种新的方法SSRMR(sparse self-representation with manifold regularization),并设计了对应的算法.

SSRMR同时利用了$l_{2,1-2}$稀疏约束和数据的流形结构信息,在获得稀疏解的同时,更加充分利用数据的局部结构信息来提高准确率,选出最具有代表性的特征.此外,为了提升算法鲁棒性,所提方法对原始数据中的异常值进行了识别处理,以获得更好的结果.本文的主要工作:①提出了一个基于SR模型的UFS方法SSRMR.其使用$l_{2,1-2}$范数进行稀疏约束,并使用流形正则化保留原始样本结构以提升准确率;②对原始数据中的异常值进行识别处理;③设计了对应的优化迭代算法计算目标函数.

1　相关工作

1.1　特征自表示

对象具有自相似性的特点,即SR属性,比如数据的每个特征都可以由其他特征表示[13].设$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_d\}$是特征集合,并且假设特征$\boldsymbol{f}_i$和$\boldsymbol{f}_j$是相互依赖的且可以表示为其他特征的线性组合.在此基础上,SR模型的定义如下:

$$\left.\begin{aligned}\boldsymbol{f}_1&=w_{11}\boldsymbol{f}_1+w_{21}\boldsymbol{f}_2+\cdots+w_{d1}\boldsymbol{f}_d+\varepsilon_1,\\ \boldsymbol{f}_2&=w_{12}\boldsymbol{f}_1+w_{22}\boldsymbol{f}_2+\cdots+w_{d2}\boldsymbol{f}_d+\varepsilon_2,\\ &\vdots\\ \boldsymbol{f}_d&=w_{1d}\boldsymbol{f}_1+w_{2d}\boldsymbol{f}_2+\cdots+w_{dd}\boldsymbol{f}_d+\varepsilon_d.\end{aligned}\right\}\tag{1}$$

其中:w_{ij}是特征表示系数;$\varepsilon_j\in\mathbf{R}^n$是误差.

1.2　流形正则化

局部几何结构可以通过流形学习获得,它在无监督学习中起着至关重要的作用[10].一般来说,基于某种相似度量的无向图可以代表数据流形结构,图的顶点代表实例,边的权重是成对样本之间的相似度.可以预见的是,特征选择后的相似样本也应该有相似的对应数据.在数学上,流形正则化的应用可以表示为

$$\arg\min_{\boldsymbol{W}}\sum_{i,j=1}^{n}S_{ij}\|\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{x}_j\boldsymbol{W}\|_2^2.\tag{2}$$

其中:$\boldsymbol{W}$为特征系数矩阵;S_{ij}是$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的相似度,可以根据各种评价指标获得.常用的相似性指标有0~1加权、高斯核、余弦相似度.通过式(2)可以看出,如果高维空间中的实例$\boldsymbol{x}_i$和$\boldsymbol{x}_j$是相似的,$\|\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{x}_j\boldsymbol{W}\|_2^2$作为嵌入样本$\boldsymbol{x}_i\boldsymbol{W}$和$\boldsymbol{x}_j\boldsymbol{W}$之间的距离应该很小,那么相似度$S_{ij}$的值就应该很大.这与期望的保留局部流形结构的策略是一致的.

1.3　无监督特征选择算法

谱特征选择SPEC(spectrum decomposition)[4]是基于图谱理论提出的一个统一的特征选择框架,可同时用于有监督和无监督特征选择,该框架首先根据数据的相似集构造其图表示,然后基于构建的图谱来评估特征;拉普拉斯分数LS(Laplacian score)[2]是SPEC的一个特例,它倾向于选择类中的特征;多聚类特征选择MCFS(multi-cluster feature selection)[5]首先使用图拉普拉斯算子实现低维嵌入,然后使用Lasso回归模型找到最佳特征子集;无监督判别特征选择UDFS(unsupervised discriminative feature selection)[3]使用散度矩阵来维护数据结构,以选择最具判别力的特征子集.RSR[13]中的每个特征都表示为其

相关特征的线性组合,其目标函数的误差项和稀疏正则项都是$l_{2,1}$范数. SCFS[12]通过子空间学习得到相似度矩阵,再利用该矩阵和$l_{2,1}$范数得到稀疏变换矩阵.子空间学习也是一种SR的机制,但相较于基于特征级SR的RSR,SCFS是基于样本级的重构,且SCFS在重构的低维空间中保持了聚类相似性. NOVRSR[14]与RSR使用了一样的特征级SR,但与RSR不同的是,其误差项通过F-范数约束,其特征表示矩阵的稀疏正则采用了非凸的$l_{2,1-2}$范数.

2 SSRMR

2.1 异常值处理

为了使相似度矩阵更加准确,同时减少由粗差导致的结果偏差,基于拉依达准则,首先检查数据经过稳健标准化后的值是否超过3,根据此找到离群点.本文中,判断x_{ij}是某一变量的离群值的标准是

$$\frac{\left|x_{ij}-\mathrm{avg}(\boldsymbol{X})\right|}{1.48\times\mathrm{avg}\left(\left|\boldsymbol{x}_i-\mathrm{avg}(\boldsymbol{X})\right|\right)}>3. \tag{3}$$

其中:avg(·)表示列平均值;$\boldsymbol{X}$为数据矩阵.分母的系数1.48确保了这种稳健的尺度收敛于正态分布下的标准差[16].本文将满足上述条件的值视为缺失值,并用平均值代替,这样可以最大限度利用原始信息,同时减少异常值的影响,整体上增强了模型的鲁棒性.

2.2 目标模型

式(1)中的SR模型对应的优化问题为

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2. \tag{4}$$

其中$\boldsymbol{f}_i$的表示参数对应于$\boldsymbol{w}_i$的值.因此,可以对$\boldsymbol{W}$使用稀疏约束来提高特征选择效率.具体地说,如果$\boldsymbol{f}_i$不重要,那么其相应的权重值将减少到0或接近0,上述目标通常由稀疏正则实现.

$l_{2,1-2}$范数相比于其他范数的优越性已经在以往的工作中[15]得到了证明,其数学定义为

$$\|\cdot\|_{2,1-2}=\|\cdot\|_{2,1}-\|\cdot\|_F. \tag{5}$$

将$\|\boldsymbol{W}\|_{2,1-2}$作为稀疏约束项,优化问题(4)可以化为

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{W}\|_{2,1-2}. \tag{6}$$

其中,$\lambda_1>0$是正则化参数.式(6)中的最小化问题可以评估与$\boldsymbol{W}$参数相对应的特征的重要性.

考虑到样本局部结构对UFS结果的影响,本文决定使用流形正则.其通过线性代数变换后为

$$R_{\mathrm{L}}=\frac{1}{2}\sum_{i,j=1}^{n}S_{ij}\left\|\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{x}_j\boldsymbol{W}\right\|_2^2=$$
$$\sum_{i=1}^{n}\left(\boldsymbol{x}_i\boldsymbol{W}\right)\left(\boldsymbol{x}_i\boldsymbol{W}\right)^{\mathrm{T}}D_{ii}-\sum_{i,\ j=1}^{n}\left(\boldsymbol{x}_i\boldsymbol{W}\right)\left(\boldsymbol{x}_j\boldsymbol{W}\right)^{\mathrm{T}}S_{ij}=$$
$$\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\boldsymbol{W}\right)-\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}\boldsymbol{W}\right)=$$
$$\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right). \tag{7}$$

其中:$\boldsymbol{S}\in\mathbf{R}^{n\times n}$是基于余弦相似的相似性矩阵[11];$\boldsymbol{D}$是一个对角元素为$D_{ii}=\sum_{j=1}^{n}S_{ij}$的对角矩阵;$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$是拉普拉斯矩阵.将$R_{\mathrm{L}}$与式(4)结合得到最终的优化目标模型,其中$\lambda_2>0$,是调节参数.

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{W}\|_{2,1-2}+\lambda_2\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}). \tag{8}$$

2.3 优化算法

式(8)中的非凸目标函数很难直接求解,但代入式(5)后,可以将其转换为如下两个凸函数相减的模型:

$$\min_{\boldsymbol{W}}\left[\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{W}\|_{2,1}+\lambda_2\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right)\right]-\lambda_1\|\boldsymbol{W}\|_F. \tag{9}$$

这种凸函数相减函数(difference of convex functions, DC)可以通过收敛的凹凸过程[17](concave-convex procedure, CCCP)来求解,下面简单介绍其步骤.设$f(x)$为可导的DC函数,存在凸函数$a(x)$和$b(x)$使$f(x)=a(x)-b(x)$. 目标函数为$\min_{x\in\mathbf{R}^n}f(x)$,其中$f_i(x)\leqslant 0, i=1,\cdots,m$. 设$b(x)$是可微的,CCCP本质上是求解这样的一组迭代子问题[18]:$x^{(t+1)}\in\arg\min_x\ a(x)-\nabla b(x^{(t)})$,其中$t$为迭代次数.

参考上述求解过程,式(9)可被转换为如下一系列凸子问题:

$$\boldsymbol{W}^{(t+1)}=\arg\min_{\boldsymbol{W}^{(t)}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}^{(t)}\|_F^2+\lambda_1\|\boldsymbol{W}^{(t)}\|_{2,1}+\lambda_2\mathrm{Tr}((\boldsymbol{X}\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}^{(t)})-\lambda_1\mathrm{Tr}((\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{C}^{(t)}). \tag{10}$$

其中$\boldsymbol{C}=\nabla\|\boldsymbol{W}\|_F=\begin{cases}\|\boldsymbol{W}\|_F^{-1}\boldsymbol{W}, & \boldsymbol{W}\neq 0\\ 0, & \boldsymbol{W}=0\end{cases}$.根据式(10),需要解决的问题可简化为

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{W}\|_{2,1}-\lambda_1\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C})+\lambda_2\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right). \tag{11}$$

式(11)中的优化问题是非光滑的,基于交替乘子法(alternating direction method of multipliers, ADMM)[19]的思想提出解决该问题的算法.首先,

引入辅助变量 $\boldsymbol{V}=\boldsymbol{W}$,将问题(11)转化为如下等式约束优化问题:

$$\min_{\boldsymbol{W},\boldsymbol{V}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{V}\|_{2,1}-\lambda_1\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C})+\lambda_2\mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right).$$

$$\text{s.t.}\quad \boldsymbol{W}-\boldsymbol{V}=0. \tag{12}$$

构造式(12)的拉格朗日函数如下:

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{V})=\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{V}\|_{2,1}+\langle\boldsymbol{\Sigma},\boldsymbol{W}-\boldsymbol{V}\rangle+\frac{\lambda_3}{2}\|\boldsymbol{W}-\boldsymbol{V}\|_F^2+\lambda_2\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W})-\lambda_1\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C}). \tag{13}$$

其中 $\boldsymbol{\Sigma}\in\mathbf{R}^{d\times d}$ 是约束 $\boldsymbol{W}-\boldsymbol{V}=0$ 的拉格朗日乘子,$\lambda_3>0$ 是惩罚参数.根据ADMM的交替迭代原理,所提算法本质上由以下3组迭代组成:

$$\boldsymbol{W}^{(h+1)}=\arg\min_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W},\boldsymbol{V}^{(h+1)}\right), \tag{14a}$$

$$\boldsymbol{V}^{(h+1)}=\arg\min_{\boldsymbol{V}}\mathcal{L}\left(\boldsymbol{W}^{(h+1)},\boldsymbol{V}\right), \tag{14b}$$

$$\boldsymbol{\Sigma}^{(h+1)}=\boldsymbol{\Sigma}^{(h)}+\lambda_3\left(\boldsymbol{W}^{(h+1)}-\boldsymbol{V}^{(h+1)}\right). \tag{14c}$$

其中 h 是迭代索引.

2.3.1　更新 $\boldsymbol{W}$

将变量 $\boldsymbol{V}$ 视为常数,式(14a)去掉与 $\boldsymbol{W}$ 无关的常数项后可简化为以下问题:

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\langle\boldsymbol{\Sigma},\boldsymbol{W}-\boldsymbol{V}\rangle+\frac{\lambda_3}{2}\|\boldsymbol{W}-\boldsymbol{V}\|_F^2+\lambda_2\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W})-\lambda_1\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C}). \tag{15}$$

利用式(15)中函数对 $\boldsymbol{W}$ 一阶求导的最优解条件,即

$$-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}-\lambda_1\boldsymbol{C}+\boldsymbol{\Sigma}+\lambda_3(\boldsymbol{W}-\boldsymbol{V})+2\lambda_2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}=0. \tag{16}$$

设 $\boldsymbol{B}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda_1\boldsymbol{C}-\boldsymbol{\Sigma}+\lambda_3\boldsymbol{V}$,代入等式(16)得到

$$\boldsymbol{W}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+2\lambda_2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}+\lambda_3\boldsymbol{I}\right)^{-1}\boldsymbol{B}. \tag{17}$$

其中 $\boldsymbol{I}\in\mathbf{R}^{d\times d}$ 是单位矩阵.

2.3.2　更新 $\boldsymbol{V}$

与更新 $\boldsymbol{W}$ 同理,只留下 $\boldsymbol{V}$ 的相关项后,式(14b)被简化为

$$\min_{\boldsymbol{V}}\lambda_1\|\boldsymbol{V}\|_{2,1}+\langle\boldsymbol{\Sigma},\boldsymbol{W}-\boldsymbol{V}\rangle+\frac{\lambda_3}{2}\|\boldsymbol{W}-\boldsymbol{V}\|_F^2. \tag{18}$$

对式(18)进行数学变换后可以得到如下等价问题:

$$\min_{\boldsymbol{V}}\frac{\lambda_1}{\lambda_3}\|\boldsymbol{V}\|_{2,1}+\frac{1}{2}\left\|\boldsymbol{V}-\left(\boldsymbol{W}+\frac{\boldsymbol{\Sigma}}{\lambda_3}\right)\right\|_F^2. \tag{19}$$

令 $\boldsymbol{M}=\boldsymbol{W}+\dfrac{\boldsymbol{\Sigma}}{\lambda_3}$,问题(19)可以转化为 d 个子问题:

$$\min_{\boldsymbol{v}_i}\sum_{i=1}^{d}\left(\frac{1}{2}\|\boldsymbol{v}_i-\boldsymbol{m}_i\|_2^2+\frac{\lambda_1}{\lambda_3}\|\boldsymbol{v}_i\|_2\right). \tag{20}$$

基于已有工作[15]可知,问题(20)的最优解为

$$\boldsymbol{v}_i=\begin{cases}\left(1-\dfrac{\lambda_1}{\lambda_3\|\boldsymbol{m}_i\|_2}\right)\boldsymbol{m}_i, & \text{if } \lambda_1<\lambda_3\|\boldsymbol{m}_i\|_2;\\ 0, & \text{otherwise.}\end{cases} \tag{21}$$

2.3.3　算　法

$\boldsymbol{W}$ 和 $\boldsymbol{V}$ 每次迭代使用上述更新规则,所提算法如下.

算法1　SSRMR算法

输入:原始数据 $\boldsymbol{X}$,拉普拉斯矩阵 $\boldsymbol{L}$,调节参数 $\lambda_1>0,\ \lambda_2>0,\ \lambda_3>0$;

输出:特征表示矩阵最优解 $\boldsymbol{W}^*$;

1.初始化 $\boldsymbol{W}^0=0,\ \boldsymbol{V}^0=0,\ \boldsymbol{\Sigma}^0=0$ and $t=0$;

2.根据式(3)识别处理 $\boldsymbol{X}$ 中的异常值;

3.While 不满足迭代停止条件 do

4.　根据式(17)更新 $\boldsymbol{W}^{(t+1)}$;

5.　根据式(21)更新 $\boldsymbol{V}^{(t+1)}$;

6.　根据式(14c)更新 $\boldsymbol{\Sigma}^{(t+1)}$;

7.　$t=t+1$;

8.End while

9.返回 $\boldsymbol{W}^*$.

2.4　计算复杂度分析

设样本数为 n,每个实例的特征数为 d,迭代次数为 t.所提出的优化方法有3个子问题,即更新 $\boldsymbol{W}$,更新 $\boldsymbol{V}$ 和更新 $\boldsymbol{\Sigma}$.求解 $\boldsymbol{W}$ 主要的计算成本是求解式(17)中的 $d\times d$ 矩阵的逆,复杂度为 $O(d^3)$.求解 $\boldsymbol{V}$ 等价于求解式(20)中的 d 个子问题,每个子问题的复杂度为 $O(n)$,因此在每次迭代中求解 $\boldsymbol{V}$ 的复杂度是 $O(nd)$.很明显每次迭代求解 $\boldsymbol{\Sigma}$ 的复杂度为 $O(d^2)$.因此所提算法的计算复杂度可粗略认为是 $O(td^3)$.

3　实　　验

3.1　实验数据集

为证明SSRMR相较于一些现有算法的优越性,本文选择在9个公开可用的数据集上进行实验验证,表1展示了这些数据集的样本大小、特征维度和实际分类数.

3.2　对比算法

为了验证所提出算法的有效性,将其与基线方法和其他7种算法进行了比较.基线方法直接使用所有数据,不进行选择特征.其他对比算法为:LS[2],SPEC[4],MCFS[5],UDFS[3],RSR[13],

SCFS[12], NOVRSR[14].

表 1 实验数据集描述

Table1 Description of the experimental dataset

数据集	样本数	特征数	分类数
DBworld	64	4 702	2
PCMAC	1 943	3 289	2
TOX-171	1 715	748	4
lung	203	3 312	5
lymphoma	96	4 026	9
nci9	60	9 712	9
JAFFE	213	1 024	10
warpPIE10P	210	2 420	10
Isolet	1 560	617	26

3.3 实验设置

涉及到 k 最近邻算法的，根据先前的经验[13]将 k 设为 5. 各算法中涉及到的正则项系数调整范围为 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. 特征维度取值范围为 $\{20, 40, 60, \cdots, 200\}$. 为评估所选特征的质量，最后把不同算法选择后的特征应用到 K-means 聚类算法中，K-means 结果受初始化影响，因此将 20 次实验输出的平均值作为最终结果.

实验运行的环境是 Windows 11 上的 Python 3.9.12，计算机 CPU 为 Intel Core i5-12400F，内存 16 GB.

3.4 实验结果与分析

聚类性能通过准确率(accuracy，ACC)和归一化互信息(normal mutual information，NMI)进行评估.

图 1 和图 2 显示了 NOVRSR 和本文所提方法 SSRMR 在各个数据集上的聚类结果，X 坐标轴是所选特征数量，Y 坐标轴是聚类结果(ACC 和 NMI). 另外，为了排除参数的影响，图中所有结果对应的相关参数都设置为 1.

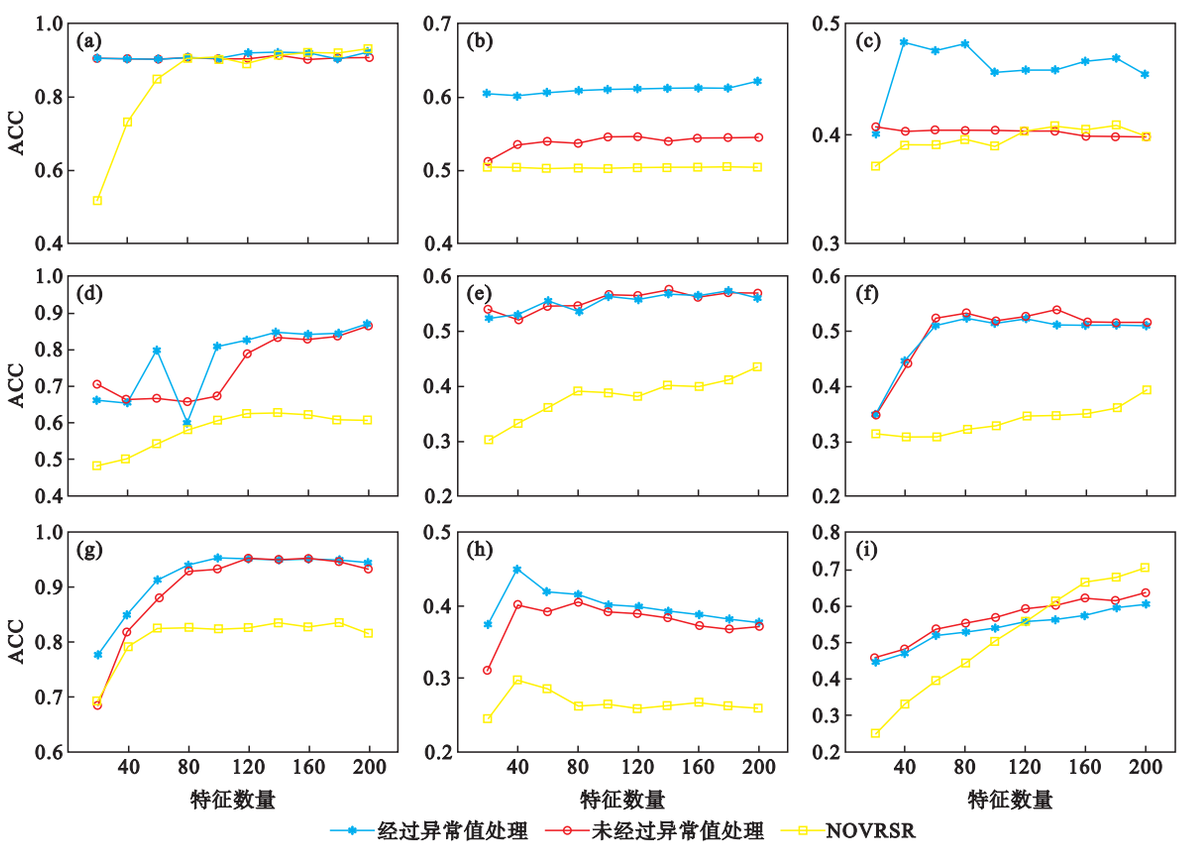

图 1 SSRMR 和 NOVRSR 的 ACC 结果比较

Fig. 1 Comparison of SSRMR and NOVRSR for ACC

(a)—DBworld; (b)—PCMAC; (c)—TOX171; (d)—lung; (e)—lymphoma; (f)—nci9; (g)—JAFFE; (h)—warpPIE10P; (i)—Isolet.

设实例数为 n. 当 $n<100$ 时，预处理的作用不大；$100<n<1\,600$ 时，预处理的贡献更为显著，但效果随着所选特征数量的增加而减少；在 $n>1\,600$ 的情况下，随着所选特征的增加，预处理始终发挥着重要作用，这是由于高维数据中所选特征的百分比相对较低，因此尚未达到预处理效果

开始下降的阶段.此外,在Isolet数据集上,预处理的结果不佳,但如果将参数设置为与最优解相对应的参数($\lambda_1=10,\lambda_2=0.1,\lambda_3=1$),如图3所示,预处理后可以得到更好的结果.总体而言,数据经过预处理后的结果更好,证明对异常数据的处理步骤是有效的.

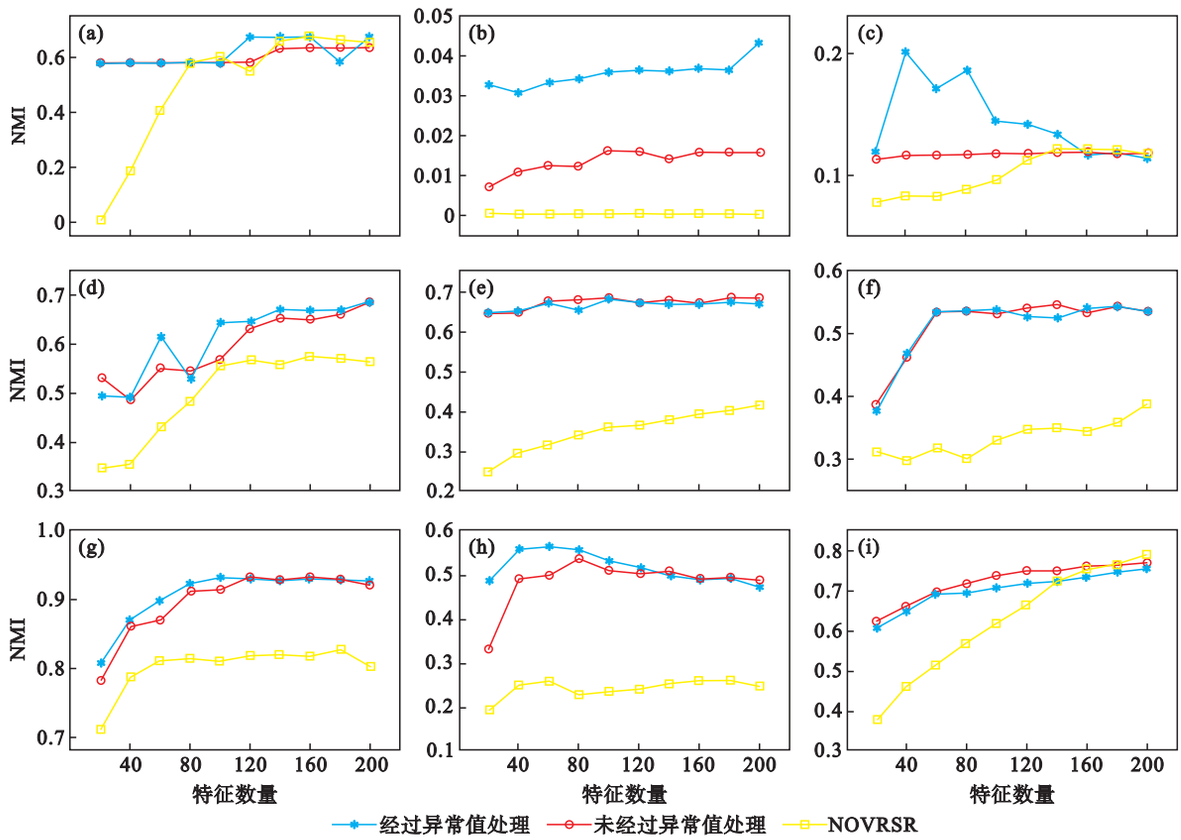

图2　SSRMR和NOVRSR的NMI结果比较

Fig. 2　Comparison of SSRMR and NOVRSR for NMI

(a)—DBworld;(b)—PCMAC;(c)—TOX171;(d)—lung;(e)—lymphoma;(f)—nci9;(g)—JAFFE;(h)—warpPIE10P;(i)—Isolet.

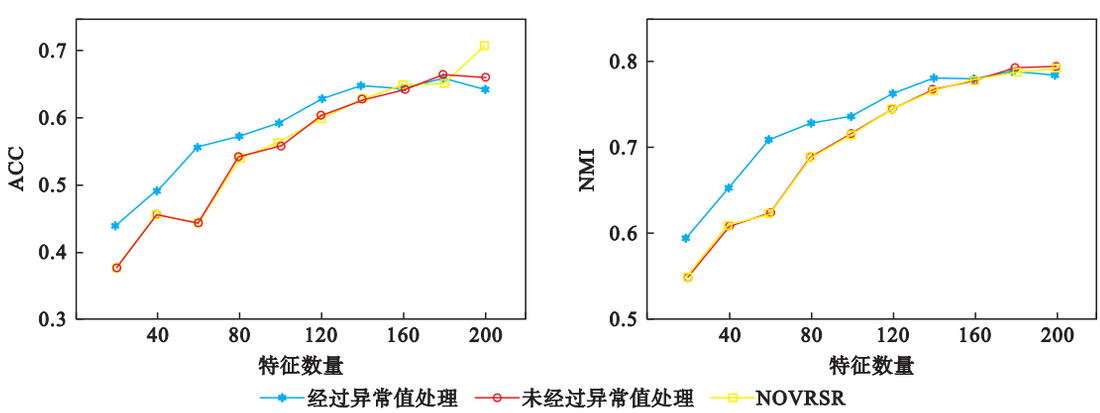

图3　Isolet数据集ACC和NMI结果

Fig. 3　Results of ACC and NMI on Isolet

从图1和图2中可以发现,SSRMR所选特征进行聚类的结果总体优于NOVRSR.在小样本数据集上选择少量特征时,两种方法的结果差异很大,NOVRSR因为能利用的数据信息有限而表现不佳,SSRMR考虑了数据的局部结构信息,获得了更好的结果,这表明通过流形正则化来保持样本间的局部结构是有助于特征选择的.

表2和表3展示了在不同数据集上所有算法在其最优参数下的聚类结果的平均值和标准差.其中最佳的结果用粗体标记,次佳的结果用下划

线标记,括号里是所选特征的数量.通过分析表2和表3,可以得出以下结论:① SSRMR总体上优于NOVRSR.验证了保存样本间的局部结构是有助于特征选择的;② 在大多数数据集中,SSRMR相较对比方法表现更好,证明了其有效性.在某些情况下,SSRMR的结果不是最佳的,但也和最佳结果接近.此外,SSRMR在某些情况下有显著的改善,例如在nci9上,其ACC和NMI分别比次佳的结果高6%左右,在warpPIE10P上的NMI比次佳的结果高11%;③ 基线方法和UFS方法的比较说明了特征选择的有效性.基线方法总体表现并不差,但它需要所有原始特征,当数据的维度很高时使用此法并不是一个明智的选择.相比之下,UFS方法只需要很少的特征就可以获得类似甚至更好的结果.

表4给出了SSRMR和另外两种UFS方法(SCFS,NOVRSR)在9个数据集上的运行时间.所提算法在部分数据集上表现较好.其中,SSRMR在TOX171数据集上花费了较多时间,是受到了参数的影响,在该数据集上实际执行的迭代次数较高.相反,在nci9花费的时间非常少,是因为实际进行的迭代次数少.

3.5 参数影响分析

固定$\lambda_3=1$,讨论参数λ_1和λ_2以及所选特征数量K对特征选择性能的影响.

从图4~图7中可以发现,在大多数情况下,随着所选特征数量K的增加,聚类表现会更好,但在达到峰值后开始下降.这种现象是因为太少的特征不能很好地代表原始特征空间,而选择太多的特征会有冗余或不相关的特征影响结果.

表2 各数据集上的聚类结果ACC

Table 2 Clustering results ACC on different datasets %

数据集	Baseline	LS	MCFS	SPEC	UDFS	RSR	SCFS	NOVRSR	SSRMR
DBworld	66.953	82.891	88.047	57.344	90.703	**93.516**	90.938	93.438	92.188
	±12.503	±4.514	±1.584	±1.116	±0.341	**±0.558**	±1.683	±0.625	±0
	(4702)	(180)	(120)	(200)	(200)	(200)	(200)	(200)	(200)
PCMAC	50.54	50.466	50.525	56.15	55.288	50.587	50.592	50.849	**62.198**
	±0.033	±0.164	±0.024	±0	±1.077	±0.022	±0	±0	**±0.104**
	(3289)	(200)	(200)	(60)	(20)	(60)	(200)	(60)	(200)
TOX171	42.69	40.029	41.316	42.456	41.345	41.55	46.345	47.222	**53.509**
	±2.661	±2.071	±0.425	±0.726	±1.373	±1.414	±3.537	±1.197	**±1.649**
	(5748)	(200)	(20)	(20)	(160)	(100)	(200)	(200)	(100)
lung	76.626	67.4631	85.813	57.512	62.192	76.502	**87.094**	78.202	86.921
	±6.922	±0.802	±1.13	±3.384	±4.975	±6.893	**±0.251**	±8.175	±1.724
	(3312)	(140)	(200)	(180)	(200)	(200)	(180)	(200)	(200)
lymphoma	59.375	53.333	61.615	46.25	58.333	54.063	**65.521**	60.365	62.76
	±4.576	±5.586	±3.642	±2.244	±5.322	±2.256	**±5.511**	±5.869	±4.669
	(4026)	(200)	(60)	(120)	(200)	(180)	(200)	(200)	(160)
nci9	43.083	40.917	43.833	40	47.333	39.417	48.5	43.833	**55.333**
	±3.13	±3.183	±3.617	±3.456	±2.759	±4.058	±2.7	±3.578	**±3.317**
	(9712)	(140)	(40)	(180)	(20)	(200)	(120)	(200)	(200)
JAFFE	85.962	70.493	82.324	66.737	80.329	85.164	94.155	88.568	**95.328**
	±2.729	±4.223	±4.839	±4.467	±5.596	±2.229	±0.909	±2.970	**±1.201**
	(1024)	(180)	(180)	(200)	(180)	(100)	(160)	(200)	(100)
warpPIE10P	26.738	35.31	44.69	30.69	41.262	30.071	26.595	46.738	**48.31**
	±2.016	±2.627	±1.912	±2.354	±1.834	±2.462	±1.349	±0.967	**±4.223**
	(2420)	(80)	(80)	(200)	(80)	(60)	(20)	(200)	(160)
Isolet	62.083	57.205	52.375	56.644	55.705	71.837	67.631	**72.356**	70.737
	±2.678	±1.759	±1.818	±2.125	±2.385	±2.309	±2.165	**±2.0**	±2.132
	(617)	(200)	(180)	(200)	(200)	(200)	(180)	(200)	(200)

表3　各数据集上的聚类结果NMI

Table 3　Clustering results NMI on different datasets　　%

数据集	Baseline	LS	MCFS	SPEC	UDFS	RSR	SCFS	NOVRSR	SSRMR
DBworld	17.375	44.727	47.326	2.479	63.825	67.671	56.353	67.671	67.671
	±15.971	±6.811	±4.008	±0	±0.882	±0	±5.407	±0	±0
	(4 702)	(120)	(120)	(40)	(200)	(160)	(200)	(160)	(200)
PCMAC	0.008	1.381	0.471	2.033	**6.926**	0.03	0.104	1.592	4.952
	±0.011	±0.236	±0	±0	**±1.382**	±0.016	±0	±0	±0
	(3 289)	(80)	(80)	(60)	(20)	(60)	(200)	(20)	(40)
TOX171	14.688	12.98	9.874	10.148	13.272	12.177	23.294	28.824	**31.578**
	±2.74	±0.921	±0.034	±0.971	±0.762	±0.987	±7.503	±1.248	**±0.512**
	(5 748)	(200)	(20)	(20)	(180)	(140)	(200)	(200)	(200)
lung	65.168	53.709	67.189	45.702	50.832	60.323	**70.218**	63.707	68.912
	±2.776	±0.9	±1.793	±4.042	±2.446	±2.805	**±0.351**	±2.07	±1.82
	(3 312)	(140)	(200)	(200)	(200)	(180)	(180)	(200)	(200)
lymphoma	69.043	58.015	71.588	47.799	65.767	66.604	70.126	69.171	**71.931**
	±3.416	±3.49	±2.466	±3.515	±4.057	±1.985	±3.126	±3.523	**±2.616**
	(4 026)	(120)	(200)	(120)	(160)	(180)	(180)	(200)	(200)
nci9	44.395	41.921	45.858	41.222	50.268	39.115	50.2	44.908	**56.22**
	±3.41	±3.811	±3.146	±2.96	±2.433	±2.84	±2.0	±3.086	**±2.949**
	(9 712)	(160)	(60)	(180)	(20)	(200)	(120)	(200)	(200)
JAFFE	86.014	69.179	85.84	69.808	85.426	83.381	92.434	88.24	**93.202**
	±1.560	±2.074	±1.601	±2.818	±1.707	±2.019	±0.614	±1.486	**±1.238**
	(1 024)	(180)	(40)	(200)	(160)	(180)	(160)	(200)	(100)
warpPIE10P	26.221	33.778	48.367	31.614	41.078	27.508	26.595	47.854	**58.906**
	±3.363	±1.906	±4.106	±1.814	±2.233	±2.584	±1.392	±1.267	**±3.268**
	(2 420)	(180)	(80)	(200)	(80)	(180)	(200)	(200)	(160)
Isolet	77.268	73.573	69.561	66.886	71.741	79.497	79.362	79.965	**80.156**
	±1.288	±0.704	±1.122	±0.703	±1.321	±1.056	±0.807	±0.891	**±1.052**
	(617)	(200)	(180)	(200)	(200)	(200)	(180)	(200)	(200)

表4　不同数据集的运行时间比较

Table 4　Running time comparison of different methods　　s

方法	DBworld	PCMAC	TOX171	lung	lymphoma	nci9	JAFFE	warpPIE10P	Isolet
SCFS	44.337	37.518	332.559	14.911	357.831	1 413.197	1.520	15.725	3.805
NOVRSR	13.204	234.375	722.079	180.455	291.301	3 033.359	11.20	80.941	2.963
SSRMR	124.265	23.767	8 707.735	80.199	10.971	148.519	2.937	3.268	45.972

显然,λ_1和λ_2的不同取值会产生不同的结果.当λ_1小于或等于1时结果更好.当λ_1太大时,会导致其调节的项(即特征自表示项)出现过拟合,影响最终结果.但总体而言,SSRMR在各数据集中对参数λ_2更敏感.这表明,由λ_2调节的项(用于保存数据的局部结构)发挥了有价值的作用.在DBworld,lymphoma和nci9中,由于它们的样本量小,可用的数据结构信息有限,其结果随着λ_2的增大整体有小幅度提升.但是,对于样本数较多的数据集,随着λ_2取值变大,ACC会出现明显的下降,这是因为λ_2过大导致的过拟合使模型性能降低.由上述分析可知,对于大样本的数据集,需要更加注意λ_2的取值.TOX171数据集是一个例外,首先,因为其特征数较少,λ_1取较大值时结果会更好;其次,TOX171受λ_1的影响明显大于λ_2,这是因为TOX171除了特征数少,分类数也少,这使得模型中每个特征对表示参数的变化会更敏感,因此受控制特征自表示项的λ_1的影响更大.相比而言,特征数同样不多的Isolet因为分类较多,数据局部结构信息的保持起到了更大的作用,因此在此数据集上λ_1对结果的影响并不明显大于λ_2.

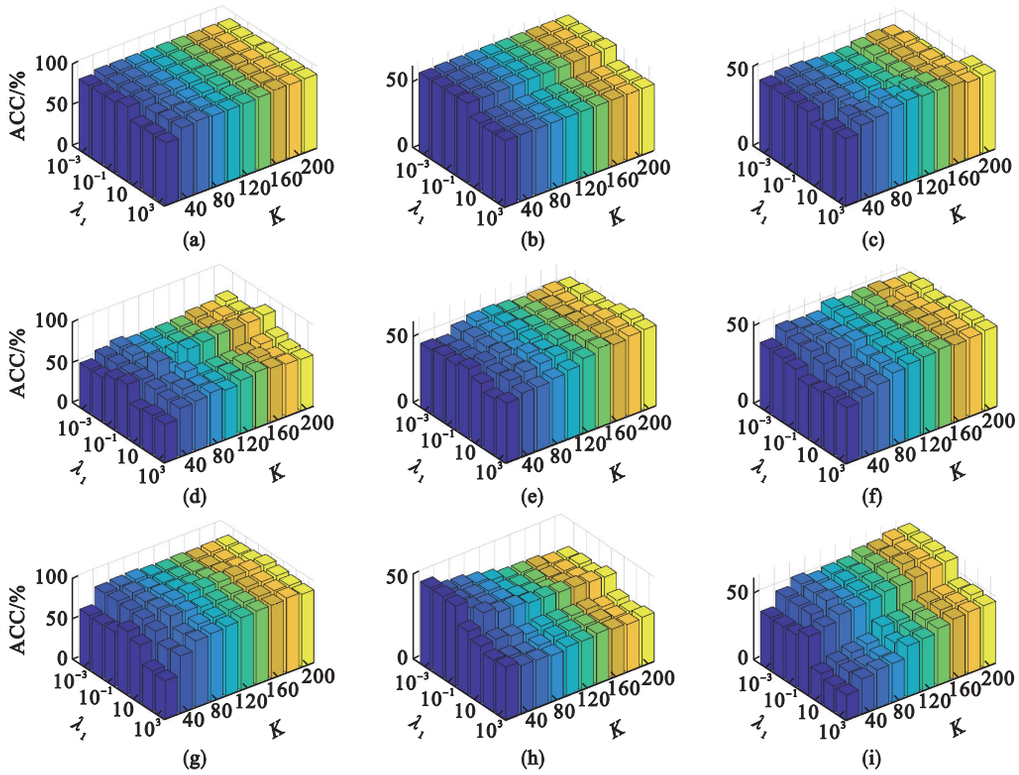

图4 λ_1和K的不同取值对ACC的影响(λ_2=1)

Fig. 4 Clustering results (ACC) with respect to λ_1 and K (λ_2=1)

(a)—DBworld; (b)—PCMAC; (c)—TOX171; (d)—lung; (e)—lymphoma; (f)—nci9; (g)—JAFFE; (h)—warpPIE10P; (i)—Isolet.

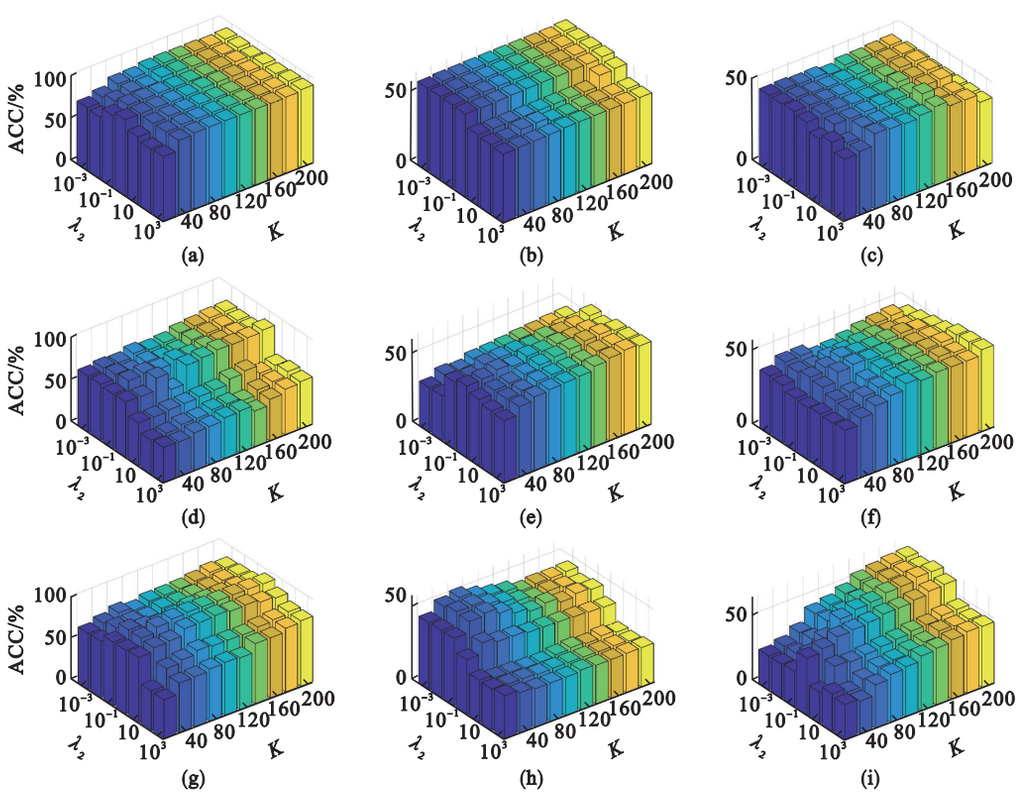

图5 λ_2和K的不同取值对ACC的影响(λ_1=1)

Fig. 5 Clustering results (ACC) with respect to λ_2 and K (λ_1=1)

(a)—DBworld; (b)—PCMAC; (c)—TOX171; (d)—lung; (e)—lymphoma; (f)—nci9; (g)—JAFFE; (h)—warpPIE10P; (i)—Isolet.

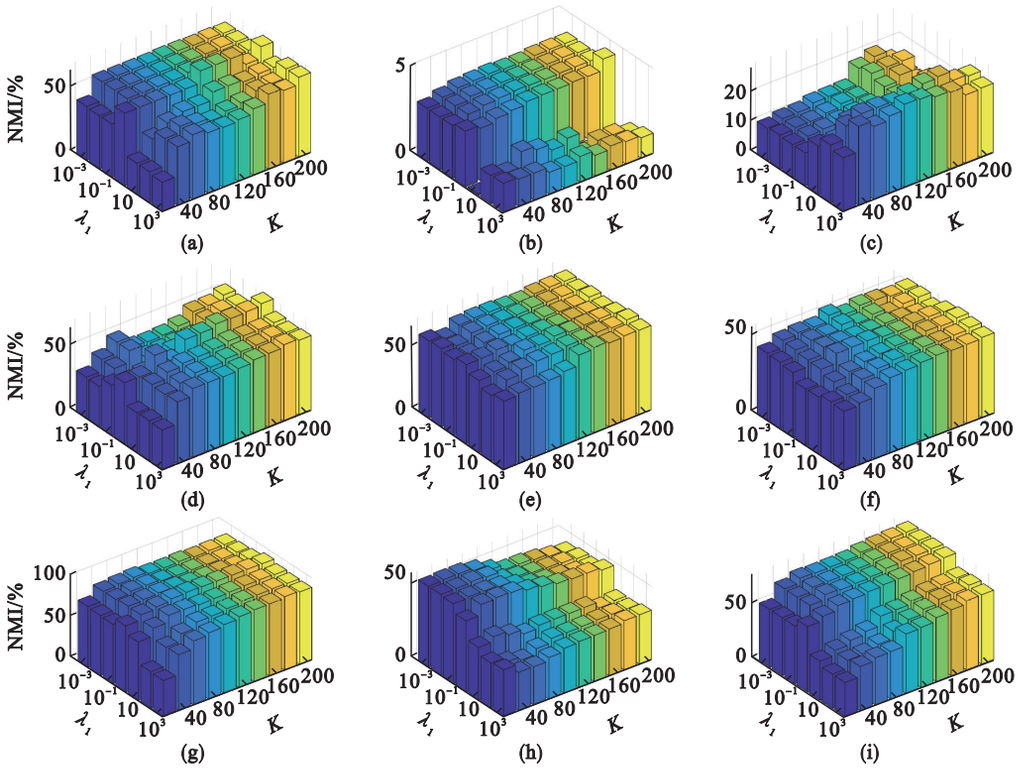

图6　λ_1和K的不同取值对NMI的影响(λ_2=1)

Fig. 6　Clustering results (NMI) with respect to λ_1 and K (λ_2=1)

(a)—DBworld; (b)—PCMAC; (c)—TOX171; (d)—lung; (e)—lymphoma; (f)—nci9; (g)—JAFFE; (h)—warpPIE10P; (i)—Isolet.

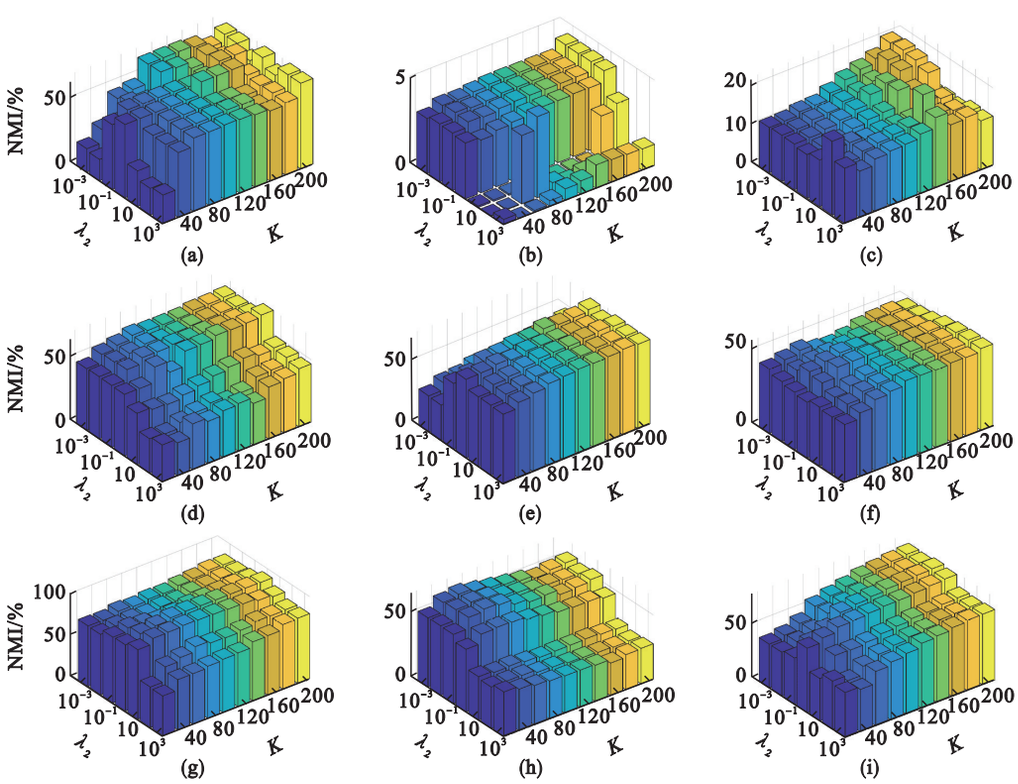

图7　λ_2和K的不同取值对NMI的影响(λ_1=1)

Fig. 7　Clustering results (NMI) with respect to λ_2 and K (λ_1=1)

(a)—DBworld; (b)—PCMAC; (c)—TOX171; (d)—lung; (e)—lymphoma; (f)—nci9; (g)—JAFFE; (h)—warpPIE10P; (i)—Isolet.

4 结 语

本文提出了一种新的基于SR模型的UFS方法.考虑到所选特征的稀疏性,使用非凸范数作为稀疏正则项.此外,为了充分利用数据的局部结构,利用了基于样本之间余弦相似性的流形正则性约束.在模型鲁棒性增强方面,采取了对原始数据进行预处理的方法来减少粗差干扰.因为目标函数很难直接求解,本文还设计了定制的迭代算法来寻找最优解.在9个数据集上将所提方法与基线方法及其他7种特征选择方法进行了比较.实验结果表明,SSRMR在各类数据集上的表现都优于其他算法.

本文的研究也有一定局限性,使用的SR模型要求每个特征都可以用其他特征的线性组合来表示,这对真实世界的数据的适用性有限,在后续研究中将寻求具有更广泛适用性的方法模型.此外,本文使用的数据都是非张量数据,为了保持最原始的多维数据结构信息,面向张量的特征选择更有优势.最后,张量数据所需的计算资源通常较大,如何使用较低的计算成本实现也是进一步研究的方向.

参考文献:

[1] Solorio-Fernández S, Ariel J, Martínez-Trinidad J. A review of unsupervised feature selection methods[J]. *The Artificial Intelligence Review*, 2020, 53(2): 907-948.

[2] He X F, Cai D, Niyogi P. Laplacian score for feature selection [C]// Proceedings of the 18th International Conference on Neural Information Processing Systems. Vancouver, 2005: 507-514.

[3] Yang Y, Shen H T, Ma Z G, et al. $L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning [C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Two. Menlo Park: AAAI Press, 2011: 1589-1594.

[4] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning [C]// Proceedings of the 24th International Conference on Machine Learning. Corvalis, 2007: 1151-1157.

[5] Cai D, Zhang C, He X F. Unsupervised feature selection for multi-cluster data [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, 2010: 333-342.

[6] 李占山,刘兆赓,俞寅,等.量子化信息素蚁群优化特征选择算法[J].东北大学学报(自然科学版),2020,41(1):17-22.
(Li Zhan-shaan, Liu Zhao-geng, Yu Yin, et al. A quantized pheromone ant colony optimization algorithm for feature selection [J]. *Journal of Northeastern University* (*Natural Science*), 2020, 41(1): 17-22.)

[7] Hu R Y, Zhu X F, Cheng D B, et al. Graph self-representation method for unsupervised feature selection[J].*Neurocomputing*, 2017, 220: 130-137.

[8] Nie F P, Zhu W, Li X L. Unsupervised feature selection with structured graph optimization [C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, 2016: 1302-1308.

[9] Lim H, Kim D. Pairwise dependence-based unsupervised feature selection[J].*Pattern Recognition*, 2021, 111: 107663.

[10] Liu X W, Wang L, Zhang J, et al. Global and local structure preservation for feature selection[J].*IEEE Transactions on Neural Networks and Learning Systems*, 2013, 25(6): 1083-1095.

[11] Li W Y, Chen H M, Li T R, et al. Unsupervised feature selection via self-paced learning and low-redundant regularization [J]. *Knowledge-Based Systems*, 2022, 240: 108150.

[12] Parsa M, Zare H, Ghatee M. Unsupervised feature selection based on adaptive similarity learning and subspace clustering [J]. *Engineering Applications of Artificial Intelligence*, 2020, 95: 103855.

[13] Zhu P F, Zuo W M, Zhang L, et al. Unsupervised feature selection by regularized self-representation [J]. *Pattern Recognition*, 2015, 48(2): 438-446.

[14] Miao J Y, Ping Y, Chen Z S, et al. Unsupervised feature selection by non-convex regularized self-representation [J]. *Expert Systems with Applications*, 2021, 173: 114643.

[15] Shi Y, Miao J Y, Wang Z Y, et al. Feature selection with $l_{2,1-2}$ regularization [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(10): 4967-4982.

[16] Bottmer L, Croux C, Wilms I. Sparse regression for large data sets with outliers[J].*European Journal of Operational Research*, 2022, 297(2): 782-794.

[17] Yuille A L, Rangarajan A. The concave-convex procedure[J]. *Neural Computation*, 2003, 15(4): 915-936.

[18] Sriperumbudur B, Lanckriet G. On the convergence of the concave-convex procedure [C]// Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, 2009: 1759-1767.

[19] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. *Foundations and Trends® in Machine Learning*, 2011, 3(1): 1-122.