

局部和稀疏保持无监督特征选择法

简彩仁, 陈晓云

(福州大学 数学与计算机科学学院, 福建 福州 350116)

摘要: 利用局部保持投影和稀疏保持投影来刻画数据的本质结构, 结合 $L_{2,1}$ 范数的组稀疏性来选择特征, 提出一种新的针对高维小样本数据集的无监督特征选择算法. 实验表明: 局部和稀疏保持无监督特征选择法是一种有效的无监督特征选择方法; 平衡参数对实验结果有较大的影响.

关键词: 局部保持投影; 稀疏保持投影; 高维小样本; 无监督; 特征选择; 聚类

中图分类号: TP 311; TP 371 **文献标志码:** A

数据维数灾难普遍存在于模式识别的许多应用中. 高维数据集不仅限制传统模式识别方法的应用, 还会显著地增加内存和时间开销. 特征选择是解决这些问题的有效手段之一^[1]. 特征选择旨在选择一些相关的特征代表原始的高维数据, 而剔除一些不相关的特征. 基于聚类的特征选择法^[2], 利用聚类算法将数据聚类, 用得到的类别信息指导特征选择. 然而, 由此获得的判别信息是不可靠的. 近年来, 随着流形学习的兴起, 学者提出了新的无监督特征选择方法, 如拉普拉斯得分^[3]、多簇特征选择方法^[4]等. 利用 $L_{2,1}$ 范数的组稀疏性, 学者提出了许多嵌入型的特征选择方法, 如稀疏限制的无监督最大化边缘的特征选择法^[5]、局部和相似保持嵌入特征选择法^[6]. 这些方法应用在高维小样本数据时, 需要求解大规模的特征值问题, 不利于问题的求解. 而联合特征选择和子空间学习法^[7]、联合局部保持投影和 $L_{2,1}$ 范数构造的特征选择, 可以克服大规模特征值的问题. 本文提出一种基于局部保持投影和稀疏保持投影的无监督特征选择方法, 并利用 $L_{2,1}$ 范数的组稀疏性质, 通过正则化 $L_{2,1}$ 范数来选择特征.

1 相关工作

局部和稀疏保持无监督特征选择法利用局部保持投影和稀疏保持投影来刻画数据的本质结构.

1.1 局部保持投影

给定数据集 $\mathbf{X} \in \mathbf{R}^{m \times n}$, 局部保持投影 (LPP)^[8] 的目标函数定义为

$$\min \sum_{i,j} \| \mathbf{y}_i - \mathbf{y}_j \| ^2 \mathbf{W}_{i,j}.$$

(1)

式(1)中: $\mathbf{y}_i = \mathbf{V}^T \mathbf{x}_i$; $\mathbf{W} = \mathbf{W}_{i,j}$ 为相似矩阵.

最小化目标函数可以使降维后的样本保持原空间的距离. 常见的相似矩阵定义是热核函数. 经过简单的代数运算, LPP 求解如下优化问题, 即

$$\begin{aligned} \min_{\mathbf{V}} \quad & \text{Tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V}), \\ \text{s. t.} \quad & \mathbf{V}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{V} = \mathbf{I}. \end{aligned}$$

(2)

式(2)中: \mathbf{V} 为投影矩阵; \mathbf{D} 为对角矩阵, $D_{i,i} = \sum_j \mathbf{W}_{i,j}$; \mathbf{L} 为图拉普拉斯矩阵, $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

1.2 稀疏保持投影

稀疏保持投影 (SPP)^[9] 用样本稀疏重构每一个样本 $\mathbf{x}_i \in \mathbf{X}$, 得到求解稀疏表示系数的模型为

$$\left. \begin{aligned} &\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_1, \\ &\text{s. t.} \quad \mathbf{x}_i = \mathbf{X}\mathbf{s}_i, 1 = \sum_i \mathbf{s}_i. \end{aligned} \right\} \tag{3}$$

式(3)中: $\|\cdot\|_1$ 为 1-范数; $\mathbf{s}_i=[s_{i,1},\cdots,s_{i,n}]$, $s_{i,j}$ 反映了 \mathbf{x}_i 和 \mathbf{x}_j 之间的关系. 因此, 将 $\mathbf{S}=(s_{i,j})_{n\times n}$ 视为仿射权矩阵是合理的. SPP 旨在寻找保持稀疏关系的投影, SPP 的目标函数为

$$\min_{\mathbf{V}} \sum_{i=1}^n \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{X} \mathbf{s}_i\|^2 \text{Tr}(\mathbf{V}^T \mathbf{X}(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{V}). \tag{4}$$

式(4)中: \mathbf{V} 为投影矩阵. 为避免平凡解, 通常引入正交约束 $\mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V}$, \mathbf{V} 可以通过求解广义特征值问题 $\mathbf{X}(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{v} = \lambda \mathbf{X} \mathbf{X}^T \mathbf{v}$ 得到.

2 局部和稀疏保持投影特征选择

2.1 目标函数

将式(2)的局部保持项和式(4)的稀疏保持项相结合, 得到目标函数为

$$\min_{\mathbf{V}} \alpha \text{Tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V}) + (1 - \alpha) \text{Tr}(\mathbf{V}^T \mathbf{X}(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{V}). \tag{5}$$

引入 $L_{2,1}$ 范数来选择有利于保持局部性和稀疏性的相关特征, 且为避免平凡解, 引入正交约束 $\mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V}$, 得到局部和稀疏保持投影特征选择模型, 即

$$\left. \begin{aligned} &\min_{\mathbf{V}} \alpha \text{Tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V}) + (1 - \alpha) \text{Tr}(\mathbf{V}^T \mathbf{X}(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{V}) + \lambda \|\mathbf{V}\|_{2,1}, \\ &\text{s. t.} \quad \mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V} = \mathbf{I}. \end{aligned} \right\} \tag{6}$$

式(6)中: α 为平衡参数, 用于平衡局部性和稀疏性; λ 为正则参数; $\|\mathbf{V}\|_{2,1}$ 定义为 $\sum_{i=1}^m (\sum_{j=1}^d |V_{i,j}|^2)^{1/2}$. 当获得投影矩阵 \mathbf{V} 后, 可以利用 \mathbf{v}_i 的 2 范数, 即 $\|\mathbf{v}_i\|_2$ 来选择特征, 其值越大表示该特征越重要.

2.2 模型求解

式(6)可以写为

$$\left. \begin{aligned} &\min_{\mathbf{V}} \text{Tr}(\mathbf{V}^T \mathbf{A} \mathbf{X} \mathbf{X}^T \mathbf{V}) + \lambda \|\mathbf{V}\|_{2,1}, \\ &\text{s. t.} \quad \mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V} = \mathbf{I}. \end{aligned} \right\} \tag{7}$$

式(7)中: $\mathbf{A} = \alpha \mathbf{L} + (1 - \alpha)(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S})$. 式(7)可以利用广义特征值问题求解. 但是, 当 \mathbf{X} 是高维小样本数据时, 式(7)需要求解大规模的特征值问题, 且会造成矩阵不可逆的问题.

类似于文献[7], 采用分步求解的方法避免上述困难. 令 $\mathbf{Y} = \mathbf{X}^T \mathbf{V}$, 有

$$\left. \begin{aligned} &\min_{\mathbf{Y}} \text{Tr}(\mathbf{Y}^T \mathbf{A} \mathbf{Y}), \\ &\text{s. t.} \quad \mathbf{Y}^T \mathbf{Y} = \mathbf{I}. \end{aligned} \right\} \tag{8}$$

式(8)的解为 $\mathbf{Y}^*=[\mathbf{y}_1,\cdots,\mathbf{y}_d]$, 其为 \mathbf{A} 的最小 d 个特征值对应的特征向量. 求解如下的问题得到式(7)的解, 即

$$\left. \begin{aligned} &\min_{\mathbf{V}} \|\mathbf{V}\|_{2,1}, \\ &\text{s. t.} \quad \mathbf{Y}^* = \mathbf{X}^T \mathbf{V}. \end{aligned} \right\} \tag{9}$$

该问题可用拉格朗日乘子法迭代求解^[7]. 拉格朗日函数为

$$L(\mathbf{V}) = \|\mathbf{V}\|_{2,1} - \text{Tr}(\mathbf{\Gamma}^T (\mathbf{X}^T \mathbf{V} - \mathbf{Y}^*)). \tag{10}$$

对 \mathbf{V} 求导得

$$\frac{\partial L(\mathbf{V})}{\partial \mathbf{V}} = 2\mathbf{D}\mathbf{V} - \mathbf{X}\mathbf{\Gamma} = 0. \tag{11}$$

式(11)中: $D_{i,i}=1/(2\|\mathbf{v}_i\|_2)$, $\mathbf{v}_i \neq 0$, 当 $D_{i,i}=0$ 时, 用一个较小的正数替代, 确保 \mathbf{D} 可逆^[8]. 不难求得

$$\mathbf{\Gamma} = 2(\mathbf{X}^T \mathbf{D}^{-1} \mathbf{X})^{-1} \mathbf{Y}^*, \tag{12}$$

将式(12)代入式(11), 可得

$$\mathbf{V} = \mathbf{D}^{-1} \mathbf{X}(\mathbf{X}^T \mathbf{D}^{-1} \mathbf{X})^{-1} \mathbf{Y}^*. \tag{13}$$

由式(12),(13)交替迭代直至收敛,可得投影矩阵 \mathbf{V} . 通过上述讨论可以得到局部和稀疏保持投影无监督特征选择法(LSP). Input:数据矩阵 \mathbf{X} ;Output:特征子集. 1) 计算拉普拉斯矩阵 \mathbf{L} 和稀疏表示矩阵 \mathbf{S} ;2) 通过式(8)计算最小 d 个广义特征值对应的特征向量 \mathbf{Y}^* ;3) 求解式(9)得到投影矩阵 \mathbf{V} ;4) 将 $\|\mathbf{v}_i\|_2$ 降序排列,选取前 p 个特征构成特征子集.

3 实验分析

相似矩阵通过 $\mathbf{W}=(|\mathbf{S}|+|\mathbf{S}^T|)/2$ 计算. 其中 \mathbf{S} 为稀疏矩阵,可以避免近邻数量的选择;平衡参数 $\alpha=0.8$. 选用数据方差(DV)、拉普拉斯得分(LS)^[3]、多簇特征选择方法(MCFS)^[4]、联合特征选择和子空间学习法(JFSSL)^[7]作为对比方法. LS,MCFS 和 JFSSL 的近邻数量取 5,MCFS,JFSSL 和 LSP 的降维维数 d 取类别个数. 通过对选取的特征子集进行聚类分析,对比聚类准确率(ACC)来验证特征选择的有效性. 实验环境为 Windows 7 系统,内存为 2 G,用 Matlab 2010b 编程实现.

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,则聚类准确率^[10]为

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

(14)

式(14)中: n 为样本总数; $\delta(x,y)$ 为函,当 $x=y$ 时,其值为 1,否则,为 0; $\text{map}(r_i)$ 为正交函数,将每一个类标签 r_i 映射成与样本自带的类标签等价的类标签.

3.1 数据集

选用 6 个公开数据集进行实验,如表 1 所示. 表 1 中:DL-BCL,LUNGANCER,LEUKEMIA,TOX 为基因表达数据集;ORL,PIE 为图像数据集.

3.2 实验结果与分析

应用每种方法选取特征子集,选取的特征个数依次设为 $\{5, 10, 15, \dots, 95, 100\}$,采用 K-means 对选取的特征子集进行聚类分析,运行 20 次. 各种方法的平均聚类准确率,如表 2 所示. 聚类准确率与特征数量(n)的关系,如图 1 所示.

表 1 数据集描述

Tab. 1 Summary of data sets

数据集	样本	特征数	类别
DLBCL	77	5 469	2
LUNGANCER	203	12 600	5
LEUKEMIA	72	5 327	3
TOX	171	5 748	4
ORL	100	10 000	10
PIE	100	10 000	10

表 2 平均聚类准确率

Tab. 2 Average clustering accuracy

%

数据集	ACC					
	LSP	JFSSL	MCFS	DV	LS	ALL
DLBCL	75.32	60.78	64.22	68.83	65.78	68.83
LUNGANCER	80.74	64.70	63.47	50.02	51.18	52.71
LEUKEMIA	76.04	66.88	62.78	56.25	47.15	62.50
TOX	45.09	45.85	40.85	40.03	40.64	40.94
ORL	74.65	72.95	68.85	50.20	66.45	67.00
PIE	49.40	38.21	42.83	25.52	26.26	30.00

由表 2,图 1 可知:局部和稀疏保持投影无监督特征选择法具有良好的特征选择能力,除 TOX 数据集外,其聚类准确率的平均值最高. 与 MCFS 和 JFSSL 相比,LSP 的聚类效果更为理想,这说明稀疏保持性质也可以刻画数据的本质结构. 与 DV 和 LS 相比,因为 DV 和 LS 只考虑独立的计算每个特征的得分,而忽略了特征之间的相互作用,所以考虑特征之间的关系可以提高聚类准确率. 此外,用 LSP 进行特征选择与保留全部特征(ALL)可以明显地提高聚类的准确率. 因此,利用局部和稀疏保持投影构造的无监督特征选择法是有效的.

3.3 参数讨论

平衡参数 α 在 $\{0, 0.1, 0.2, \dots, 0.9, 1.0\}$ 变化时,平均聚类准确率的情况,如图 2 所示. 由图 2 可知:总体上,平衡参数 α 对 LSP 的影响是明显的;当 α 为 $0.6 \sim 0.9$ 时,LSP 的聚类准确率在较高的水平上保持相对稳定. 在这一范围内稀疏保持项的比重较大,说明稀疏保持项可以提高特征选择的能力.

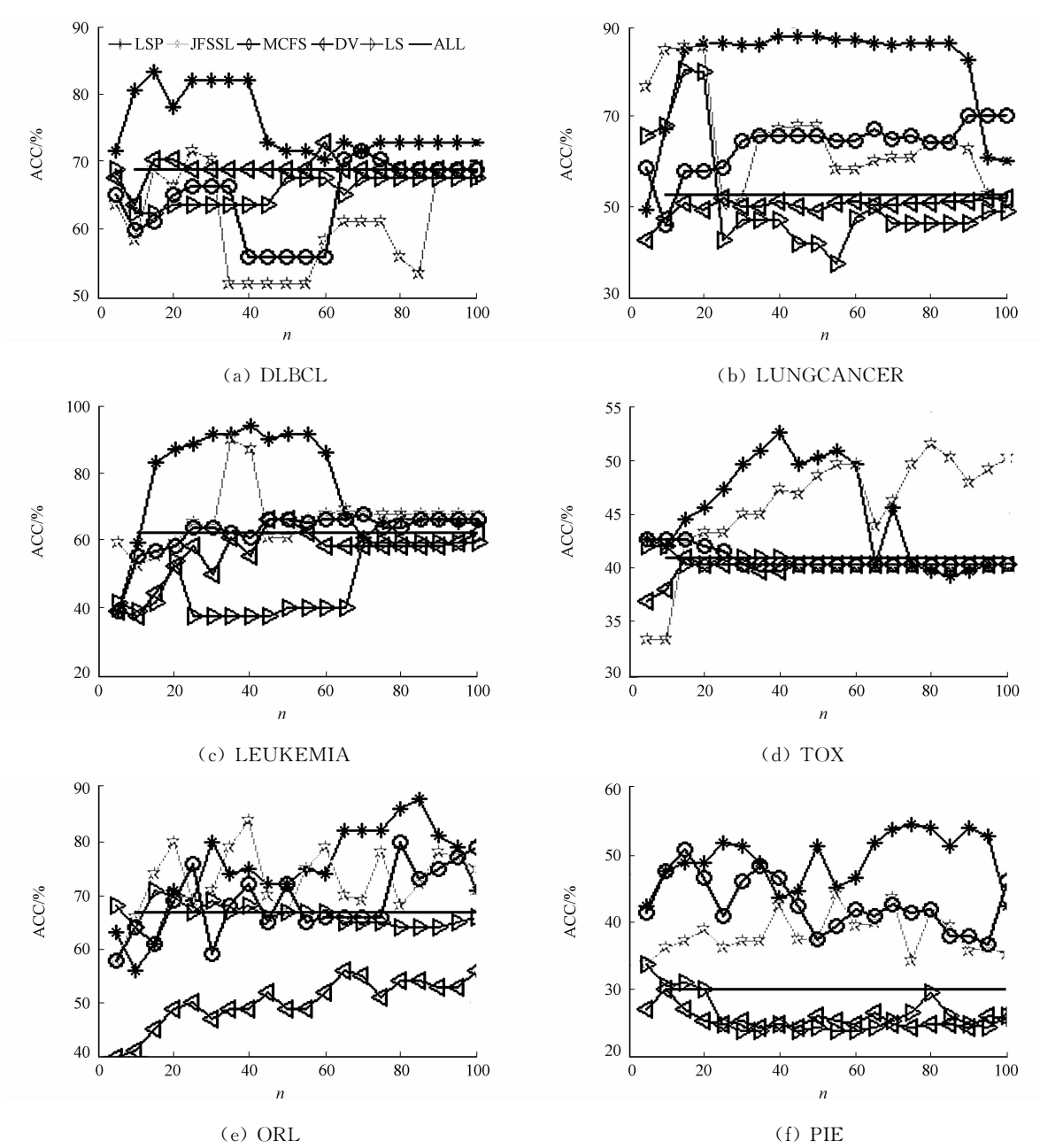
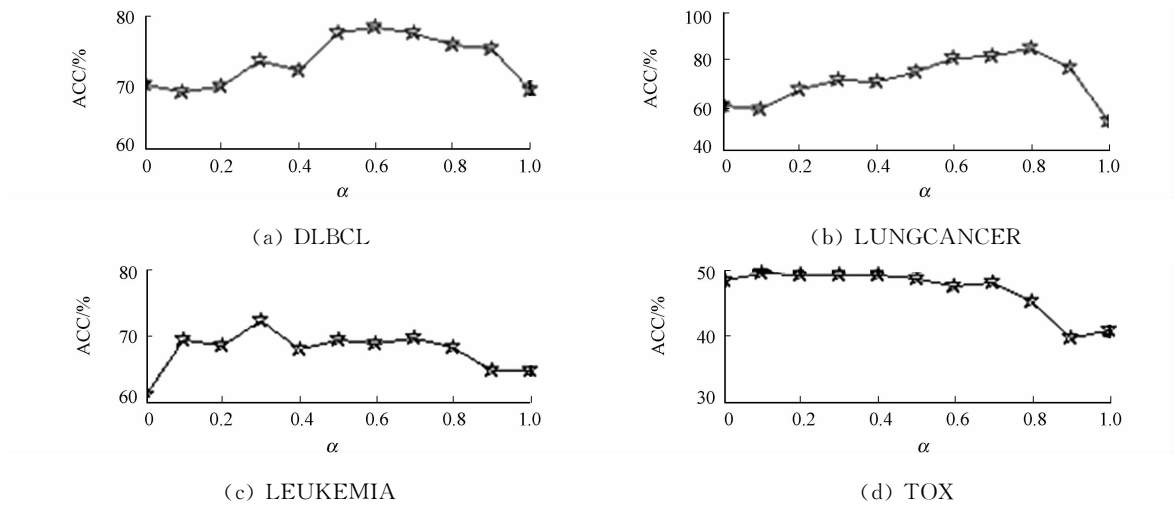


图 1 聚类准确率与选择的特征数量关系

Fig. 1 Relation between clustering accuracy and the number of selected features



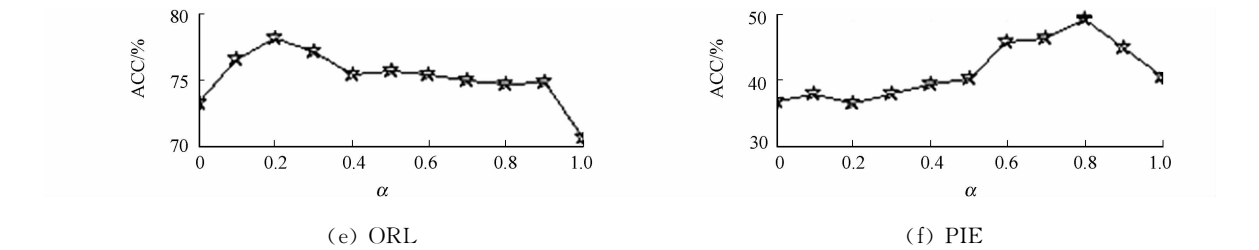


图 2 不同 α 的聚类准确率
Fig. 2 Clustering accuracy of different α

4 结束语

提出局部和稀疏保持无监督特征选择法,利用局部保持投影和稀疏保持投影来刻画数据的本质结构,利用 $L_{2,1}$ 范数的组稀疏性来筛选特征.实验结果表明:LSP 是一种有效的无监督特征选择方法.LSP 方法的平衡参数 α 对实验结果有较大的影响,如何自适应地选取该参数将在今后的研究中给出.

参考文献:

[1] 徐峻岭,周毓明,陈林,等.基于互信息的无监督特征选择[J].计算机研究与发展,2012,49(2):372-382.

[2] 张莉,孙钢,郭军.基于 K-均值聚类的无监督的特征选择方法[J].计算机应用研究,2005,22(3):23-24.

[3] HE Xiao-fei,CAI Deng,NIYOGI P. Laplacian score for feature selection[C]// Advances in Neural Information Processing Systems. Vancouver:[s. n.],2005:507-514.

[4] CAI Deng,ZHANG Chi-yuan,HE Xiao-fei. Unsupervised feature selection for multi-cluster data[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC: ACM,2010:333-342.

[5] YANG Shi-zhun,HOU Chen-ping,NIE Fei-ping,et al. Unsupervised maximum margin feature selection via $L_{2,1}$ -norm minimization[J]. Neural Computing and Applications,2012,21(7):1791-1799.

[6] FANG Xiao-zhao,XU Yong,LI Xue-long,et al. Locality and similarity preserving embedding for feature selection [J]. Neurocomputing,2014,128:304-315.

[7] GU Quan-quan,LI Zhen-hui,HAN Jia-wei. Joint feature selection and subspace learning[C]// The 22nd International Joint Conference on Artificial Intelligence. Barcelona:[s. n.],2011:1294-1299.

[8] HE Xiao-hui,NIYOGI P. Locality preserving projections[C]// Proceedings of the 17th Annual Conference on Neural Information Processing Systems. Columbia:[s. n.],2003:153-160.

[9] QIAO Li-shan,CHEN Song-can,TAN Xiao-yang. Sparsity preserving projections with applications to face recognition[J]. Pattern Recognition,2010,43(1):331-341.

[10] CAI Deng,HE Xiao-fei,WU Xiao-yun,et al. Non-negative matrix factorization on manifold[C]// Proceedings of International Conference on Data Mining. Pisa:IEEE Press,2008:63-72.

Unsupervised Feature Selection Using Locality and Sparsity Preserving

JIAN Cai-ren, CHEN Xiao-yun

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: By locality preserving projection and sparsity preserving projection to represent the intrinsic geometrical structure of the data set and use the group sparse of $L_{2,1}$ norm, one new unsupervised feature selection method for high-dimensionality small sample data set is proposed. Experimental results show that the method is effective and sensitive to balance parameter.

Keywords: locality preserving projection; sparsity preserving projection; high-dimensionality small sample; unsupervised; feature selection; clustering