计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.20190800095

改进的局部和相似性保持特征选择算法

李金霞[1]　赵志刚[1]　李　强[1]　吕慧显[2]　李明生[1]
1 青岛大学计算机科学技术学院　山东 青岛 266071
2 青岛大学自动化与电气工程学院　山东 青岛 266071
(ljx7130@163.com)

摘　要　LSPE(Locality and Similarity Preserving Embedding)特征选择算法首先基于KNN定义图结构来保持数据的局部性,再基于定义图学习数据的低维重构系数来保持数据的局部性和相似性。两个步骤独立进行,缺乏交互。由于近邻个数是人为定义的,使得学习到的图结构不具备自适应的近邻,不是最优的,进而影响算法性能。为优化LSPE算法的性能,提出改进的局部和相似性保持特征选择算法,将图学习与稀疏重构、特征选择并入同一个框架,使得图学习和稀疏编码同时进行,其要求编码过程是稀疏的,自适应近邻的和非负的。所提算法旨在寻找一个能保持数据的局部性和相似性的投影,并对投影矩阵施加$l_{2,1}$范数,进而选择能够保持局部性和相似性的相关特征。实验结果表明,改进后的算法减少了主观人为影响,消除了选择特征的不稳定性,对数据噪声鲁棒性更强,提高了图像分类的准确率。

关键词:稀疏重构;局部和相似性保持;特征选择;无监督学习
中图法分类号　TP391.4

Improved Locality and Similarity Preserving Feature Selection Algorithm

LI Jin-xia[1],ZHAO Zhi-gang[1],LI Qiang[1],LV Hui-xian[2] and LI Ming-sheng[1]
1 College of Computer Science and Technology,Qingdao University,Qingdao,Shandong 266071,China
2 College of Automation and Electrical Engineering,Qingdao University,Qingdao,Shandong 266071,China

Abstract　LSPE (Locality and similarity preserving embedding) feature selection algorithm firstly maintains the locality of the data based on the pre-defined graph structure of the KNN,and then maintains the locality and similarity of the data based on the low-dimensional reconstruction coefficients that define the learning data of the graph. The two steps are independent and lack of interaction. Since the number of nearest neighbors is artificially defined,the learned graph structure does not have adaptive nearest neighbors and is not optimal,which will affect the performance of the algorithm. In order to optimize the performance of LSPE,an improved locality and similarity preserving feature selection algorithm is proposed. The proposed algorithm incorporates graph learning,sparse reconstruction and feature selection into the same framework,making graph learning and sparse coding are carried out simultaneously. The coding process is required to to be sparse,adaptive neighbor and non-negative. The goal is to find a projection that can maintain the locality and similarity of the data,and apply a $l_{2,1}$-norm to the projection matrix,and then select the relevant features that can maintain locality and similarity. Experimental results show that the improved algorithm reduces the subjective influence,eliminates the instability of selecting features,is more robust to data noise,and improves the accuracy of image classification.

Keywords　Sparse reconstruction,Locality and similarity preserving,Feature selection,Unsupervised learning

1　引言

随着大数据时代的到来,数据中含有大量冗余信息,噪声信息的高维数据呈爆发式增长,对后续的数据处理和分析产生了严重影响[1]。为从大量高维数据中得到最有效信息,需要对数据进行降维处理。特征选择旨在根据一定的准则从原始特征空间中选择最具代表性的特征子集,进而降低数据维度。

近年来,基于谱图理论的特征选择方法受到了学者们的广泛关注。Zhao等[2]依据k近邻准则建立权重矩阵和图拉普拉斯矩阵。Liu等[3]进一步提出了以迹比准则为评价机制的半监督特征选择算法TRCFS(TraceRatio Criterion for Feature Selection)。谱图计算简单,但构建过程往往需要占用较大的时间和空间资源。Chang等[4]提出多标签的半监督特征选择方法CSFS(Convex Semi-supervised Multi-label Feature Selection),无须构建谱图,利用全局线性回归函数进行特征选择,节约了时间和空间成本,但忽略了真实数据的底层流形结构。而局部线性嵌入(Locality linear embedding,LLE)[5]和拉普拉斯特征映射(Laplacianeigenmaps,LE)[6]能够较准确的发现数据的内在流形结构。但文献[7]指出:大多数流形学习方法都不能处理新的样本。因此,He等[8]提出局部保持投影(Locality Preserving Projection,LPP),使用学习到的投影矩阵将新数据转换到低维子空间,可以有效处理新的样本数据。He等[9]进一步提出了近邻保持嵌入(NPE)用于保持数据的局部近邻结构。

基金项目:国家重点研发项目(2017YFB0203102)
This work was supported by the National Key Research and Development Program of China (2017YFB0203102).
通信作者:赵志刚(zhaolhx@163.com)

目前，基于稀疏子空间学习的特征选择算法相继被提出，如稀疏保持投影（Sparsity Preserving Projections，SPP）[10]和局部坐标编码（Local Coordinate Coding，LCC）[11]。但这些方法大多直接在原始数据表示上进行稀疏编码，忽略了原始数据中存在的冗余信息和噪声，影响了编码质量。Fang 等[12]提出基于稀疏约束的局部和相似性保持方法（LSPE），将特征选择和稀疏编码并入到一个框架中，解决了直接在原始数据上编码的问题。但 LSPE 算法存在以下问题：1）使用两个独立的步骤，首先基于 KNN 预先定义图结构，再基于定义图学习投影矩阵。由于这两个独立分开的学习步骤不能交互，因此算法难以达到最优。2）最近邻准则不能得到丰富的判别信息。3）近邻参数和核函数的参数需要人为定义，带有强烈的主观意愿，尽管可以交叉验证，但却非常耗时。本文结合 SPP 算法和 LCC 算法，将图学习与稀疏重构、特征选择并入同一个框架，改进了 LSPE 算法两步走策略，使得图学习和特征选择学习同时进行，确保学习到最优的图结构，减少人为干扰；引入 $l_{2,1}$ 范数约束，使得编码过程是稀疏的、自适应近邻和非负的。同时，将学习到的数据低秩稀疏表示矩阵嵌入在降维过程中，从而获得足够的判别信息，使得算法对数据噪声鲁棒性更强，保证特征选择算法的有效性。此外，该模型非光滑，因此提出了一种有效的迭代算法来优化模型。

2 相关工作

2.1 $l_{2,1}$ 范数

给定矩阵 $\boldsymbol{M}\in R^{a\times b}$，其 $l_{2,1}$ 范数定义为：

$$\| \boldsymbol{M} \|_{2,1}=\sum_{i=1}^{a}\sqrt{\sum_{j=1}^{b}\boldsymbol{M}_{ij}^{2}} \tag{1}$$

2.2 稀疏保持投影

SPP 用较少样本重构每一个样本，获取稀疏重构向量 $\boldsymbol{s}_i$ 的模型为：

$$\min_{\boldsymbol{s}_i} \| \boldsymbol{s}_i \|_1 \quad \text{s.t. } \boldsymbol{x}_i=\boldsymbol{X}\boldsymbol{x}_i,\mathbf{1}=\mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i \tag{2}$$

其中，$\| \cdot \|_1$ 表示 l_1 范数，$\mathbf{1}\in R^n$ 是一个元素全为 1 的向量。

稀疏的重构矩阵 $\boldsymbol{S}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n]$ 中的每个元素 s_{ij} 能准确反映样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的近邻关系，因此可用作仿射权重矩阵。SPP 通过学习投影矩阵 $\boldsymbol{Q}$ 来保持样本之间的稀疏重构关系：

$$\min_{\boldsymbol{Q}}\sum_{i=1}^{n} \| \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{s}_i \|^2 \tag{3}$$

式（3）可转化为如下目标函数：

$$\min_{\boldsymbol{Q}}\sum_{i=1}^{n} \| \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{s}_i \|^2=\boldsymbol{Q}^{\mathrm{T}}\Big(\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{s}_i)(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{s}_i)^{\mathrm{T}}\Big)\boldsymbol{Q} \tag{4}$$

可通过求解如下特征方程得到最优 $\boldsymbol{Q}$：

$$\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{q}_i=\lambda\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{q}_i \tag{5}$$

2.3 局部的坐标编码

文献[13]指出，一定假设下数据的局部性比稀疏性更重要。且为保证良好的分类效果，相似的数据应具有相似的编码，如：若 x_i，x_j 两个样本在几何分布上十分接近，那其最优重构系数也应十分接近。LCC 旨在通过编码近邻样本来线性重构每个样本，即保证近邻样本获得相似的编码系数。

LCC 按照式（6）计算重构系数 $s_i\in\boldsymbol{R}^K$：

$$\min_{\boldsymbol{s}_i} \| \boldsymbol{x}_i-\boldsymbol{D}\boldsymbol{s}_i \|_2^2+\lambda\sum_{k=1}^{K}|S_i| \| \boldsymbol{D}_{*k}-\boldsymbol{x}_i \|_2^2 \tag{6}$$

其中，$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_K]$ 表示字典，$\boldsymbol{D}_{*k}$ 表示其第 k 列。正则化项 $\lambda\sum_{k=1}^{K}|\boldsymbol{s}_i| \| \boldsymbol{D}_{*k}-\boldsymbol{x}_i \|_2^2$ 保证每个输入样本都能被字典中的近邻样本线性重构。

3 本文方法

3.1 目标函数

式（6）中的正则化项可通过计算样本和字典原子间的距离，自动确定近邻样本，并为其分配较大权重，保证了学习到的图具有自适应的近邻。为将数据的局部性和相似性嵌入在低维空间中，并能同时优化稀疏重构矩阵 $\boldsymbol{S}$ 和投影矩阵 $\boldsymbol{A}$，定义目标函数为：

$$\begin{aligned}&\min_{\boldsymbol{A},\boldsymbol{S}}\sum_{i=1}^{n} \| \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{s}_i) \|^2+\beta\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]\\&\text{s.t. } \boldsymbol{S}\geqslant 0,\boldsymbol{S}_{ii}=0,\forall i\end{aligned} \tag{7}$$

其中，$\boldsymbol{A}\in\boldsymbol{R}^{m\times d}$ 为投影矩阵，d 表示低维空间的维度。$\boldsymbol{M}\in\boldsymbol{R}^{n\times n}$，$\boldsymbol{M}_{ij}= \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_2^2$。$\boldsymbol{E}\in\boldsymbol{R}^{n\times n}$ 为元素全为 1 的矩阵。$\odot$ 表示矩阵之间的哈达玛运算。

式（7）中的正则项 $\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]$ 具有以下作用[14]：1）保证每一个样本都能准确地被其近邻样本线性重构，且其近邻样本占有较大权重；2）可保证相似的样本具有相似的编码系数；3）容许编码过程是稀疏的、自适应近邻的和非负的。

算法目的是寻找一个能够保持数据局部性和相似性的投影，则要求投影矩阵具有行一致性稀疏的属性，即引入 $l_{2,1}$ 范数约束投影矩阵 $\boldsymbol{A}$。令 $\boldsymbol{A}_{i*}$ 表示矩阵 $\boldsymbol{A}$ 的第 i 行，用来度量第 i 个特征的重要性。最终的目标函数为：

$$\begin{aligned}&\min_{\boldsymbol{A},\boldsymbol{S}}\sum_{i=1}^{n} \| \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{s}_i) \|^2+\beta\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]+\alpha \| \boldsymbol{A} \|_{2,1}\\&\text{s.t. } \boldsymbol{S}\geqslant 0,\boldsymbol{S}_{ii}=0,\forall i\end{aligned} \tag{8}$$

其中，$\alpha\geqslant 0$，$\beta\geqslant 0$ 表示对应的参数。

3.2 优化算法

由于所提目标函数非光滑，难以直接优化。我们提出一个迭代优化算法迭代更新稀疏重构矩阵 $\boldsymbol{S}$ 和投影矩阵 $\boldsymbol{A}$。为方便优化，目标函数重写为：

$$\begin{aligned}&\min_{\boldsymbol{A},\boldsymbol{S}}\sum_{i=1}^{n} \| \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{s}_i) \|^2+\beta\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]+\alpha \| \boldsymbol{A} \|_{2,1}\\&=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})+\alpha \| \boldsymbol{A} \|_{2,1}+\\&\quad\beta\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]\end{aligned} \tag{9}$$

若固定 $\boldsymbol{S}$，定义函数：

$$\boldsymbol{L}(\boldsymbol{A})=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})+\alpha \| \boldsymbol{A} \|_{2,1} \tag{10}$$

其中，$\boldsymbol{K}=\boldsymbol{I}-\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}$

定义对角阵 $\boldsymbol{U}\in\boldsymbol{R}^{m\times m}$ 如下：

$$\boldsymbol{U}_{ii}=\frac{1}{2 \| \boldsymbol{A}_{i*} \|_2} \tag{11}$$

$\boldsymbol{L}(\boldsymbol{A})$ 可重写为：

$$\boldsymbol{L}(\boldsymbol{A})=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})+\alpha\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{A}) \tag{12}$$

为得到问题的解，要求投影矩阵 $\boldsymbol{A}$ 正交，即有如下目标函数：

$$\begin{aligned}&\arg\min_{\boldsymbol{A}}\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^{\mathrm{T}}+\alpha\boldsymbol{U})\boldsymbol{A})\\&\text{s.t. } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}\end{aligned} \tag{13}$$

投影矩阵 $\boldsymbol{A}$ 可进行如下的特征值分解：

$$(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^{\mathrm{T}}+\alpha\boldsymbol{U})\cdot\boldsymbol{a}_i=\lambda\boldsymbol{a}_i \tag{14}$$

假设 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_d]$ 是问题（14）的解，那么最小的 d 个特征值对应的向量 $\boldsymbol{a}_i(i=1,2,\cdots,d)$ 即为问题的解。

若固定 $\boldsymbol{A}$，则有如下目标函数：

$$\begin{aligned}&\boldsymbol{C}(\boldsymbol{S})=\min_{\boldsymbol{S}}\mathrm{tr}(\boldsymbol{D}(\boldsymbol{I}-\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})\boldsymbol{D}^{\mathrm{T}})+\beta\mathrm{tr}[\boldsymbol{E}(\boldsymbol{S}\odot\boldsymbol{M})]\\&\text{s.t. } \boldsymbol{S}\geqslant 0,\boldsymbol{S}_{ii}=0,\forall i\end{aligned} \tag{15}$$

其中，$\boldsymbol{D}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$。式（15）是个有约束的优化问题，式（15）可分

解为 n 个独立的非负权重稀疏编码问题：

$$\min_{\boldsymbol{S}}\sum_{i}^{n}\sum_{k=1}^{n}\beta\boldsymbol{M}_{*i}^{k}\boldsymbol{S}_{*i}^{k}+\mathrm{tr}(\boldsymbol{D}(\boldsymbol{I}-\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})\boldsymbol{D}^{\mathrm{T}}) \quad (16)$$

$$\text{s.t. } \boldsymbol{S}\geqslant 0, \boldsymbol{S}_{ii}=0, \forall i$$

其中，$\boldsymbol{S}_{*i}^{k}$ 表示 $\boldsymbol{S}_{*i}$ 的第 k 个成分，$\boldsymbol{M}_{*i}^{k}$ 表示 $\boldsymbol{M}$ 的第 i 个列向量的第 k 个成分。我们采用交替方向法（Alternating Direction Method，ADM）[15]来求解式(16)。

得到最优的 $\boldsymbol{S}$ 和 $\boldsymbol{A}$ 后，计算矩阵 $\boldsymbol{A}$ 每一行对应的 l_2 范数，对其降序排列，选择前 h 个对应的特征组成最优的特征子集。

以上优化过程可概括为算法 1。

算法 1

输入：训练数据集 $\mathbf{X}\in\mathbf{R}^{m\times n}$；正则化参数 α，β；初始化 $\mathbf{U}\in\mathbf{R}^{m\times m}$ 为单位矩阵；初始化 $\mathbf{S}=1_{n\times n}$，其中 $1_{n\times n}$ 为元素全为 1 的矩阵

输出：投影矩阵 $\mathbf{A}\in\mathbf{R}^{m\times d}$，$\mathbf{S}$

1. 设 $\mathbf{t}=1$，计算 $\mathbf{K}=\mathbf{I}-\mathbf{S}-\mathbf{S}^{\mathrm{T}}+\mathbf{S}^{\mathrm{T}}\mathbf{S}$

2. 计算 $\mathbf{P}_t=\mathbf{X}\mathbf{K}\mathbf{X}^{\mathrm{T}}+\alpha\mathbf{U}$

3. 计算 $\mathbf{A}_t=[\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_d]$，$\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_d$ 为 $\mathbf{P}_t$ 最小的 d 个特征值对应的特征向量

4. 更新 $\mathbf{U}_{t+1}=\begin{bmatrix}\frac{1}{2\|\mathbf{A}_1^t\|_2} & & 0\\ & \ddots & \\ 0 & & \frac{1}{2\|\mathbf{A}_m^t\|_2}\end{bmatrix}$

5. 更新 $\mathbf{S}$：求解式(16)

6. t=t+1，转至第 1 步直到收敛

7. Return $\mathbf{A}$，$\mathbf{S}$

4 实验及结果分析

4.1 对比方法及实验数据

实验采用的对比方法如下。

1）All-feature。保留原始数据全部特征，分类结果用作本实验基准线。

2）Fisher[16]。依据 Fisher 准则独立评价每个特征重要性的经典特征选择方法。

3）FSNM（Feature Selection via Joint $l_{2,1}$ Norms Minimization）[17]。引入 $l_{2,1}$ 范数模型进行特征选择。

4）TRCFS（TraceRatio Criterion for Feature Selection）。通过引入迹比准则来减少噪声干扰。

5）CSFS（Convex Semi-supervised Multi-label Feature Selection）。一种基于全局线性约束的快速半监督特征选择方法。

6）LSPE（Locality and Similarity Preserving Embedding）。基于稀疏约束的局部和相似性保持的无监督特征选择算法。

将各算法应用到多种开源数据库，包括 BREAST，WINE，UMIST，ORL，YALE[18]。其中 UMIST，ORL，YALE 是人脸数据集。表 1 列出所选数据集的相关描述。

表 1 实验数据集

Table 1 Experimental datasets

Data	Size	Dimension	Class	Selected feature
BREAST	699	10	2	{1,2,…,9}
WINE	178	13	3	{1,2,…,12}
UMIST	575	644	20	{100,200,…,600}
ORL	400	1024	40	{100,200,…,900}
YALE	165	1024	15	{100,200,…,900}

对于每种方法涉及到的参数，参照文献[19]将参数范围设定为$\{10^{-2},1,10^{2},10^{4},10^{6},10^{8}\}$。每组实验独立重复 5 次，并记录 5 次结果的平均值。每组实验中，对于每个数据集均随机选择 10%，20%，40%样本作为训练集，剩余样本作为测试集。另，对半监督方法保留训练样本标签，无监督方法删除训练样本标签。实验在 Matlab2018 环境中进行，选择 LIBSVM 作为分类工具，其中 LIBSVM 的最优参数采用 5-fold 交叉验证。

4.2 分类准确度分析

各种方法在 5 种数据集上的分类准确度如表 2 所列。

表 2 各种方法在 5 种数据集上的分类准确度对比

Tabel 2 Classification accuracy comparison of different feature selection algorithms on 5 datasets

（单位：%）

Labeled percentage	Data	All-feature	Fisher	FSNM	TRCFS	CSFS	LSPE	Ours
10%	BREAST	64.60	94.66	94.82	94.71	94.86	93.68	**94.89**
	WINE	78.63	82.21	86.30	85.21	85.34	83.65	**86.54**
	UMIST	67.76	68.56	70.79	70.44	70.50	70.83	**73.24**
	ORL	43.94	41.56	42.19	41.67	44.22	46.61	**51.67**
	YALE	26.98	26.65	27.72	27.65	28.05	41.93	**42.72**
20%	BREAST	65.14	95.50	95.78	95.82	95.79	94.21	**96.31**
	WINE	88.95	93.28	92.44	92.86	93.99	91.06	**93.62**
	UMIST	89.65	89.17	87.30	88.87	90.57	86.77	**89.65**
	ORL	58.50	58.75	59.75	58.75	59.68	64.94	**65.07**
	YALE	43.30	42.76	42.30	42.45	43.18	48.67	**50.00**
40%	BREAST	65.76	96.33	96.43	96.33	96.67	95.80	**96.58**
	WINE	91.40	94.95	95.33	94.95	95.33	93.52	**94.67**
	UMIST	97.04	97.33	97.10	97.33	97.44	94.82	**97.34**
	ORL	81.25	81.58	81.33	81.41	83.50	83.83	**87.50**
	YALE	55.96	55.99	56.40	56.16	56.16	55.56	**57.56**

观察表 2 实验结果可知，本文算法在 5 种数据集上均具有较高的分类准确率，尤其在人脸数据集上表现突出，提高了图像分类的准确率，有效改进了 LSPE 算法。

同时通过对比发现，随着有标签样本的比例增加，半监督算法如 CSFS 算法在某些数据集上的准确率比本算法略高，表明了标签信息对提高分类准确率的积极作用。在准确率相差不大的情况下，特征选择方法通过选择最具有判别性的特征子集，有效减少了原始数据中的冗余信息和噪声信息干扰，大幅度降低了数据维度，节省了后续数据处理分析的时间、空间成本。

4.3 收敛性分析

为证明本文提出的模型的收敛性，对 BREAST，WINE，YALE，ORL 4 个数据集进行了收敛性分析，如图 1 所示，其中 α，β 值均设为 1。

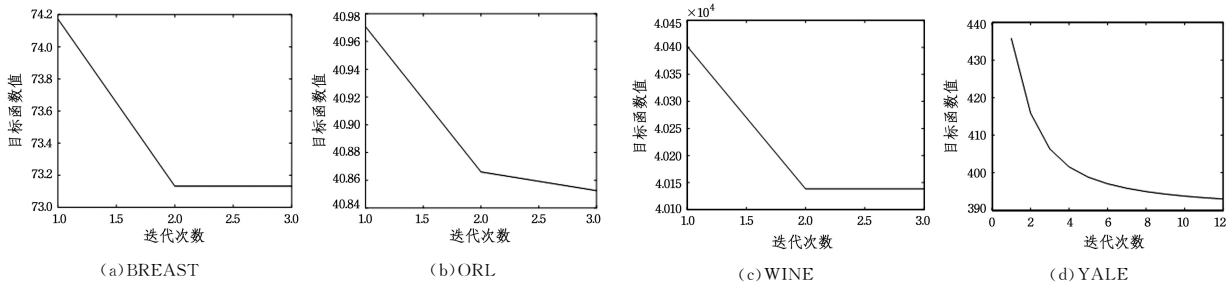

图 1 本文方法收敛性分析

Fig. 1 Convergence analysis of proposed method

从图 1 可以看出，不同数据集经过不同的迭代次数达到收敛状态，各数据集迭代次数均不超过 15，实验数据表明了本文提出的方法具有较高的执行效率。

4.4 参数敏感性分析

在 BREAST，YALE，UMIST，ORL 4 组数据集上进行实验，分析所提模型的 α，β 参数。由图 2 可以看出，α 和 β 的选择依赖于数据集的选择，样本关于参数 α，β 的分类性能随数据集的不同而变化。在对数据集 BREAST，UMIST，ORL 的实验中，数据的分类准确率关于参数 α，β 和特征个数都较为稳定。而 YALE 数据集关于参数 α，β 和特征个数都不稳定，但 $\alpha=1$ 比 $\beta=1$ 时的分类准确率表现较稳定。这说明 $\alpha=1$ 对大部分数据集来说都较为稳定。

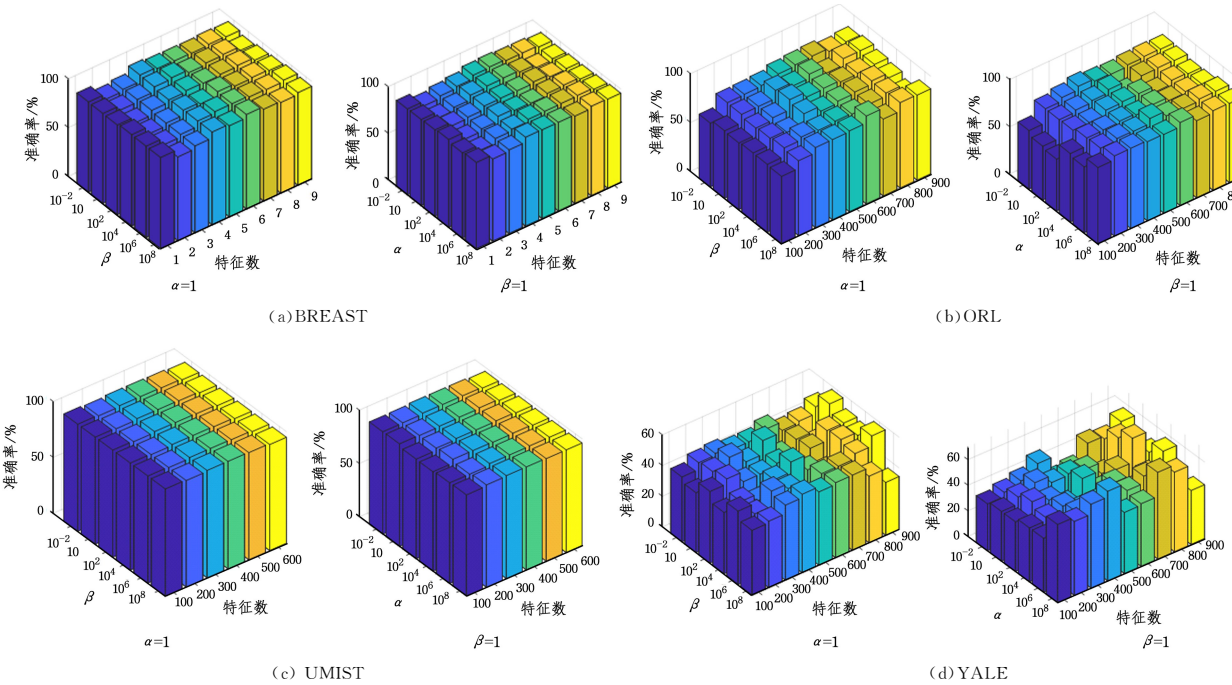

图 2 参数敏感性分析

Fig. 2 Parameter sensitivity analysis

结束语 本文将图学习、特征选择和稀疏编码并入到一个框架中，同时学习稀疏重构和投影矩阵，容许编码过程是稀疏的、自适应近邻的和非负的。将学习到的数据低秩稀疏表示矩阵嵌入降维过程中，从而获得足够的判别信息，使得算法对数据噪声鲁棒性更强，保证特征选择算法的有效性。实验数据证明了本文算法在人脸数据集上优势更明显，在与半监督方法的对比中，亦证明了标签信息对提高准确率的积极作用。为有效地利用实际应用中的少标样数据，未来研究方向为尝试引入标签信息，将模型拓展为半监督特征选择方法。

参 考 文 献

[1] LI T，MENG Z，NI B，et al. Robust Geometric ℓp-norm Feature Pooling for Image Classification and Action Recognition[J]. Image & Vision Computing，2016，55(P2)：64-76.

[2] ZHAO Z，LIU H. Semi-supervised feature selection via spectral analysis.[C]// Proceedings of the 2007 SIAM International Conference on Data Mining. Minneapolis，Minnesota：SIAM，2007：26-28.

[3] LIU Y，NIE F，WU J，et al. Efficient semi-supervised feature selection with noise insensitive trace ratio criterion[J]. Neurocomputing，2013，105(Complete)：12-18.

[4] CHANG X，NIE F P，YANG Y，et al. A convex formulation for semi-supervised multi-label feature selection[C]// Proc of the 28th AAAI Conference on Artificial Intelligence. 2014：1171-1177.

[5] ROWEIS S T. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science，2000，290(5500)：2323-2326.

[6] BELKIN M，NIYOGI P. LaplacianEigenmaps for Dimensionality Reduction and Data Representation[M]. MIT Press，2003.

[7] BENGIO Y，VINCENT P. Out-of-Sample Extensions for LLE，Isomap，MDS，Eigenmaps，and Spectral Clustering[C]// International Conference on Neural Information Processing Systems. MIT Press，2003.

[8] HE X, YAN S, HU Y, et al. Face Recognition Using Laplacian Faces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328-340.
[9] HE X. Neighborhood preserving embedding [C]// Tenth IEEE International Conference on Computer Vision. 2005: 1208-1213.
[10] QIAO L, CHEN S, TAN X. Sparsity preserving projections with applications to face recognition[J]. Pattern Recognition, 2010, 43(1): 331-341.
[11] KAI Y, TONG Z, GONG Y. Nonlinear Learning Using Local Coordinate Coding[C]// Advances in Neural Information Processing Systems. 2009: 2223-2231.
[12] FANG X, XU Y, LI X, et al. Locality and similarity preserving embedding for feature selection[J]. Neurocomputing, 2014, 128: 304-315.
[13] WANG J, YANG J, KAI Y, et al. Locality-constrained Linear Coding for image classification[C]// Computer Vision & Pattern Recognition. 2010: 3360-3367.
[14] FANG X, XU Y, LI X, et al. Learning a Nonnegative Sparse Graph for Linear Regression[J]. IEEE Transactions on Image Processing, 2015, 24(9): 2760-2771
[15] YANG J, ZHANG Y. Alternating Direction Algorithms for L1-Problems in Compressive Sensing[J]. SIAM Journal on Scientific Computing, 2011, 33(1): 250-278.
[16] FUKUNAGA, KEINOSUKE. Introduction to statistical pattern recognition[M]. Academic Press, 1972.
[17] NIE F, HUANG H, CAI X, et al. Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization[C]// Advances in Neural Information Processing Systems. 2010: 1813-1821.
[18] HAN Y, XU Z, MA Z, et al. Image classification with manifold learning for out-of-sample data[J]. Signal Processing, 2013, 93(8): 2169-2177.
[19] YAN F, WANG X D. A semi-supervised feature selection method based on local discriminant constraint[J]. Pattern Recognition AndArtificial Intelligence, 2017, 30(1): 89-95.

LI Jin-xia, born in 1994, postgraduate. Her main research interests include machine learning and so on.

ZHAO Zhi-gang, born in 1973, professor, is a member of China Computer Federation. His main research interests include image processing, machine learning and compressed sensing.

(上接第 479 页)

对 HBase 配置参数敏感即可。自动参数调优算法是一种有效的工具，可以自动选择最优的不同配置参数组合，而且可以帮助用户在吞吐量和延迟之间做出良好的权衡。

参 考 文 献

[1] COOPER B F, SILBERSTEIN A, TAM E, et al. Benchmarking cloud serving systems with YCSB[C]// Proc. 1st ACM Symp. Cloud Comput. (SoCC), New York, NY, USA, 2010: 143-154.
[2] HBase at Taobao, accessed on May 26, 2017. [OL]. http://www. eygle. com/digest/2012/03/hbase-at-taobao. html.
[3] Apache HBase Team. Apache HBase Reference Guide[OL]. http://hbase. apache. org/book. html.
[4] BAO X, LIU L, XIAO N, et al. Policy-driven configuration management for NoSQL[C]// Proc. IEEE 8th Int. Conf. Cloud Comput. . 2015: 245-252.
[5] BREIMAN L. Bagging predictors [J]. Mach. Learn. , 1996, 24(2): 123-140.
[6] 赵宏，张洁，侯鲁健，等. 并行 GA_ANN 预测模型研究[J]. 计算机工程与应用，2011(22).
[7] COOPER B F, SILBERSTEIN A, TAM E, et al. Benchmarking-cloud serving systems with YCSB[C]// Proc. 1st ACM Symp. Cloud Comput. (SoCC), New York, NY, USA, 2010: 143-154.
[8] BRODER A, MITZENMACHER M. Network applications of bloom fifilters: A survey[J]. Internet Math. , 2004, 1(4): 485-509.
[9] BREIMAN L. Random forests[J]. Mach. Learn. , 2001, 45(1): 5-32.
[10] EFRON B, TIBSHIRANI R J. An Introduction to Bootstrap [M]. Boca Raton, FL, USA: CRC Press, 1994.
[11] LIAW A, WIENER M. lassifification and regression by randomforest[J]. R News, 2002, 2(3): 18-22.
[12] COOPER B F, et al. PNUTS: Yahoo!'s hosted data serving platform[J]. J. Proc. VLDB Endowment, 2008, 1(2): 1277-1288.
[13] Apache Cassandra, accessed on May 26[OL]. http://incubator. apache. org/cassandra/.
[14] CALDER B, et al. Windows azure storage: A highly available cloud storage service with strong consistency[C]// Proc. 23rd ACM Symp. Oper. Syst. Principles, 2011: 143-157.
[15] Apache CouchDB, accessed on May 26, 2017. [OL]. http://couchdb. apache. org/.
[16] SCIORE E. SimpleDB: A simple java-based multiuser syst for teaching database internals[J]. ACM SIGCSE Bull. , 2007, 9(1): 561-565.
[17] Project Voldemort, accessed on May 26[OL]. http://project-voldemort. com.

XU Jiang-feng, born in 1965, Ph.D, professor, is a member of China Computer Federation. His research interests include data encryption technology, and network security technology.

TAN Yu-long, born in 1994, postgraduate, is a member of China Computer Federation. His research interests include information security, network security technology, and machine lear-ning.