基于子空间学习的图稀疏属性选择算法*

钟 智[1]，何 威[2†]，程德波[2]，胡荣耀[2]，刘星毅[3]
(1. 广西师范学院 计算机与信息工程学院，南宁 530023；2. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004；3. 广西钦州学院，广西 钦州 535000)

摘 要： 针对处理高维度属性的大数据属性约减方法进行了研究，发现属性选择和子空间学习是属性约简的两种常见方法，其中属性选择具有很好的解释性，子空间学习的分类效果优于属性选择，而往往这两种方法是各自独立进行应用的。为此，综合这两种属性约简方法，设计出新的属性选择方法，即利用子空间学习的两种技术(即线性判别分析(LDA)和局部保持投影(LPP))，考虑数据的全局特性和局部特性，同时设置稀疏正则化因子实现属性选择。基于分类准确率、方差和变异系数等评价指标的实验结果表明，该算法相比其他算法，能更有效地选取判别属性，并能取得很好的分类效果。

关键词： 属性约简；属性选择；子空间学习；线性判别分析；局部保持投影；稀疏学习
中图分类号： TP181 **文献标志码：** A **文章编号：** 1001-3695(2016)09-2679-04
doi:10.3969/j.issn.1001-3695.2016.09.026

Graph sparse for feature selection algorithm based subspace learning

Zhong Zhi[1], He Wei[2†], Cheng Debo[2], Hu Rongyao[2], Liu Xingyi[3]
(1. *College of Computer & Information Engineering, Guangxi Teachers Education University, Nanning* 530023, *China*; 2. *Guangxi Key Laboratory of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin Guangxi* 541004, *China*; 3. *Qinzhou University, Qinzhou Guangxi* 535000, *China*)

Abstract: Aimed at dimensionality reduction method for processing high-dimensional features of big data to research, and found that feature selection and subspace learning are two traditional methods of dimensionality reduction. Where feature selection contains interpretable characteristics while subspace learning shows better classification performance than the former. And it often applied these two methods independently. This paper proposed a novel feature selection method by integrating subspace learning(i.e., via using LDA and LPP, respectively, for preserving the global structures and the local structure of data) with feature selection (i.e., via a sparse regularization term). Experimental results based on classification accuracy, variance and coefficient of variation as comparative evaluations, show that this algorithm compared to other algorithms, is more effectively to select discriminating property and can achieve good classification results.

Key words: dimensionality reduction; feature selection; subspace learning; LDA; LPP; sparse learning

0 引言

在计算机视觉、模式识别和生物研究等诸多领域中，大数据特征通常用高维表示。高维数据不仅增大了数据储存空间和运算时间复杂度，而且在数据处理过程中容易导致维灾难等问题[1]。因此，在高维数据中进行属性约简是分类、聚类等机器学习模型的关键步骤[2]。

属性约简的目标是从高维数据集中选择一个更为紧实准确的子集作为数据的新属性集。这样被选择出的子集维度比原始集低，使得后继知识发现过程更高效，而且冗余和噪声属性在属性约简后被去除，从而达到更准确的聚类和分类目的。总体来说，属性约简技术可以分为两类[3]：a) 属性选择，即从原始属性集中选择出一个最能表示原始特征的子集，本质是选择一种信息量大的属性子集，常见的方法有 t 检验法和稀疏正则化线性回归等；b) 子空间学习，即把原始数据集转换到一个低维度的子空间。常见的子空间学习方法有主成分分析(principal component analysis，PCA)、邻域保持嵌入(neighborhood preserving embedding，NPE)、线性判别分析(linear discriminant analysis，LDA)和局部保持投影(locality preserving projections，LPP)[3]。其中，LDA 是将高维的模式样本投影到最佳的判别矢量空间，通过类内方差和类间方差之比，反映观测结果的固有全局信息[4]；而 LPP 的主要思想是，样本之间在高维空间中距离比较近，则它们投影到低维空间距离也应该很近，由此来保持数据间局部的相似结构信息[5]。属性选择方法具有很好的解释性，但子空间学习比属性选择更有效，却因为得到的数据为原始属性的线性组合而不具有解释性[6]。因此，结合属性选择和子空间学习两种方法的优点进行属性约简显然是一种理想的方法。

收稿日期：2015-04-17；修回日期：2015-06-06 基金项目：国家自然科学基金资助项目(61170131，61263035，61363009)；国家“863”计划资助项目(2012AA011005)；国家“973”计划资助项目(2013CB329404)；广西自然科学基金资助项目(2012GXNSFGA060004)；广西八桂创新团队和广西百人计划资助项目；广西高校科学技术研究重点项目(2013ZD041)；广西研究生教育创新计划项目(YCSZ2015095，YCSZ2015096)

作者简介：钟智(1963-)，男，广西梧州人，副教授，主要研究方向为机器学习和数据挖掘；何威(1989-)，男(通信作者)，河南商丘人，硕士，主要研究方向为数据挖掘、机器学习(8328682@qq.com)；程德波(1990-)，男，江西丰城人，硕士，主要研究方向为数据挖掘、机器学习；胡荣耀(1992-)，男，江西景德镇人，硕士，主要研究方向为数据挖掘、机器学习；刘星毅，男，广西钦州人，硕士，主要研究方向为数据挖掘.

本文充分考虑了属性选择和子空间学习的优点，提出了一种新的属性约简方法，即基于子空间的图稀疏属性选择算法(graph sparse for feature selection algorithm based subspace learning，SG_FS)。本文首先利用最小二乘损失函数加上一个能导致行稀疏的 $l_{2,1}$-范数正则化惩罚项进行属性选择；接着把数据的全局信息(即通过 LDA)和局部信息(即通过 LPP)加入到已设计好的属性选择框架，目的是选取类辨别和抗噪声属性[1]。由于本文算法同时考虑数据的全局信息和局部信息(即融入 LDA 和 LPP)，它具有比单一的子空间学习(如 LDA 或 LPP)或者属性选择方法等更强的分类能力，而且本文方法具有可解释性。再者，考虑大多数的属性选择方法只专注二分类问题和实际应用中二类和多类分类问题都十分普遍等情况，本文提出的属性选择方法可以适应各种分类情形。

1 算法描述

1.1 稀疏辨别属性选择

本质上，数据的特点是由少量关键特征决定的，稀疏学习可实现特征的自动选择。在稀疏学习中，l_0-范数为最有效的稀疏正则因子。但因其求解为 NP 难题，故很多资料采用近似正则项 l_1-范数来替代 l_0-范数；而 $l_{2,1}$-范数等导致行稀疏，已经被证明比 l_1-范数更适合于属性选择[7]。

本文定义属性矩阵 $X\in \mathrm{R}^{d\times n}$，其中 d 和 n 分别是特征变量数和样本数，$Y\in \mathrm{R}^{c\times n}$ 定义类指示矩阵或类标签矩阵，其中 c 是类数。对任意矩阵根据多任务学习最小二乘回归模型，结合对重构系数作 $l_{2,1}$-范数的正则化约束，首先定义多类属性选择问题公式：

$$\min_W \frac{1}{2}\| Y - W^{\mathrm{T}}X\|_F^2 + \lambda \| W\|_{2,1} \tag{1}$$

其中：$W\in \mathrm{R}^{d\times c}$ 是回归系数矩阵，用来对 X 进行属性选择；λ 是正参数用来调节对损失项的惩罚；$\| W\|_{2,1}$ 为 $l_{2,1}$-范数。式(1)能赋值大权重给重要的属性，赋值小权重给不重要的属性，这种方法已经成功应用于二分类[8]。对于多任务学习，式(1)把各个类作为一个任务，并用此方法来定义不同类之间的联系[9]。然而，在式(1)的现有形式中，不能保证被选择属性的类辨别能力，即不具备子空间学习的特性。

为此，本文提出一种全新的区别于式(1)的属性选择方法，即在式(1)的基础上，同时考虑数据的全局分布和数据间的局部拓扑关系。a)使用线性判别分析(LDA)考虑基于类内方差和类间方差来寻找类的鉴别属性；b)使用保局投影(LPP)的概念来保留数据间的拓扑关系。

关于鉴别属性选择的 LDA 准则，可以使用一种直接的方法作为目标函数的惩罚项，其中正则项被定义为

$$R_G = \frac{W^{\mathrm{T}}\Sigma_w W}{W^{\mathrm{T}}\Sigma_b W} \tag{2}$$

其中：Σ_w 和 Σ_b 分别表示类内方差和类间方差。由于式(2)是非凸的，难以找到目标函数的最优解。幸运的是多类 LDA[10] 的提出，即可通过最大化式(2)来寻找最优子空间。式(2)能被线性回归模型等效推导，即通过在式(1)中定义类指示矩阵 $Y=[y_{i,k}]$，如下所示：

$$y_{i,k}=\begin{cases}\sqrt{\dfrac{n}{n_k}}-\sqrt{\dfrac{n_k}{n}} & \text{如果 } l(x_i)=k\\ -\sqrt{\dfrac{n_k}{n}} & \text{否则}\end{cases} \tag{3}$$

其中：$l(x_i)$ 表示属于 x_i 的类标签，n_k 是类 k 的样本大小。也就是说，利用式(3)定义的类指示矩阵 Y，能有效地保持数据的全局信息，即在原始空间里的数据分布。

对于保持数据之间的拓扑关系，即局部信息，本文使用 LPP，即通过定义如下目标函数：

$$\min_W \sum_{i,j}(W^{\mathrm{T}}x_i - W^{\mathrm{T}}x_j)^2 s_{ij} \tag{4}$$

$S=[s_{i,j}]\in \mathrm{R}^{n\times n}$ 是相似度矩阵，通过热核(heat kernel)($H(x_i,x_j)=\exp\left[-\dfrac{\| x_i-x_j\|^2}{\sigma}\right]$，其中参数 $\sigma\in \mathrm{R}^+$)来定义每对数据 x_i 和 x_j 之间的相似性 $s_{i,j}$。为使式(4)融入式(1)可推导：

$$\frac{1}{2}\sum_{i,j}(W^{\mathrm{T}}x_i - W^{\mathrm{T}}x_j)^2 s_{ij} = \sum_i (W^{\mathrm{T}}x_i d_{ij} x_i^{\mathrm{T}} W) - \sum_{ij}(W^{\mathrm{T}}x_i s_{ij} x_i^{\mathrm{T}} W) =$$
$$\mathrm{tr}(W^{\mathrm{T}}XDX^{\mathrm{T}}W) - \mathrm{tr}(W^{\mathrm{T}}XSX^{\mathrm{T}}W) = \mathrm{tr}(W^{\mathrm{T}}XLX^{\mathrm{T}}W) \tag{5}$$

其中：$L=D-S$，$D=[d_{i,i}=\sum_j s_{i,j}]\in \mathrm{R}^{n\times n}$ 是对角矩阵。

最后，本文将式(3)与(5)融入式(1)，得到 SG_FS 算法的目标函数为

$$\min_W \frac{1}{2}\| Y - W^{\mathrm{T}}X\|_F^2 + \lambda_1 \mathrm{tr}(W^{\mathrm{T}}XLX^{\mathrm{T}}W) + \lambda_2 \| W\|_{2,1} \tag{6}$$

其中，Y 在式(3)中被定义，λ_1 和 λ_2 是调和参数。λ_1 被设计为平衡 $\mathrm{tr}(W^{\mathrm{T}}XLX^{\mathrm{T}}W)$ 和 $\| Y - W^{\mathrm{T}}X\|_F^2$ 之间数量级的参数，λ_1 值越大则对于目标函数式(6)来说，LPP 的贡献就越大，反之 LPP 的贡献就越小。显然，式(6)在属性约简的框架下融合了属性选择和子空间学习的理念，而且同时考虑了两种不同且互补的子空间学习方法，即 LDA 保持数据的全局信息，LPP 保持数据的局部信息。算法的伪代码如算法 1。

SG_FS 区别于式(1)的方法在于：a)不像先前的基于稀疏线性回归的属性选择方法，本文方法使用线性判别分析(LDA)和拉普拉斯算子找到了类区分和抗噪声回归矩阵；b)相对于子空间学习方法，如 PCA、LDA 和 LPP 等都具有解释的局限性，而 SG_FS 方法是直接选择原始空间的特征，因此它具有对结果的直观考察；c)不同于传统的基于式(2)，即标准的 LDA，本文方法使用 Fisher 准则但仍然在原始属性空间内操作，因此还具有对所选属性的直观解释。此外，传统的 LDA 只能从 c 类分类任务中找到最多 $(c-1)$ 维属性，如从 3 类分类中最多能找到 2 维属性，而式(6)可从 c 类分类中选择最多的 d(通常 $d>>c$)维属性。

算法 1 本文算法伪代码

输入：训练样本，正则化参数 λ_1、λ_2。

输出：分类准确率。

1 通过式(3)得出类指示矩阵 $Y=[y_{i,k}]$；

2 依据所选择的模型：

$\min_W \frac{1}{2}\| Y - W^{\mathrm{T}}X\|_F^2 + \lambda_1 \mathrm{tr}(W^{\mathrm{T}}XLX^{\mathrm{T}}W) + \lambda_2 \| W\|_{2,1}$ 调用算法 2 求解优化问题得到回归系数矩阵 $W\in \mathrm{R}^{d\times c}$；

3 利用 W 对原始属性集 X 进行属性选择后得到的属性集作为新的属性集；

4 对新的属性集采用 SVM 分类。

1.2 优化分析求解

式(6)是一个凸且非光滑的函数，为此，在本文中提出一种新的加速邻近梯度法来求解式(6)。首先，将式(6)按引导邻近梯度方法拆分成：

$$f(W) = \frac{1}{2}\| Y - W^{\mathrm{T}}X\|_F^2 + \lambda_1 \mathrm{tr}(W^{\mathrm{T}}XLX^{\mathrm{T}}W) \tag{7}$$

$$L(W) = f(W) + \lambda_2 \| W\|_{2,1} \tag{8}$$

注意到 $f(W)$ 是凸且可微的，然而 $\| W\|_{2,1}$ 是凸的但非光

滑。为了使用邻近梯度方法来优化求解,本文使用以下优化准则迭代更新 W:

$$W(t+1) = \arg\min_{W} G_{\eta(t)}(W, W(t)) \tag{9}$$

其中: $G_{\eta(t)}(W, W(t)) = f(W(t)) + \langle \nabla f(W(t)), W - W(t) \rangle + \frac{\eta(t)}{2} \| W - W(t) \|_F^2 + \lambda_2 \| W \|_{2,1}$, $\nabla f(W(t)) = (XX^{\mathrm{T}} + \lambda_1 XLX^{\mathrm{T}}) W(t) - XY^{\mathrm{T}}$, $\eta(t)$ 和 $W(t)$ 分别是调节参数和从 t 迭代获得的 W 的值。

在式(9)中忽略不依赖 W 的项,可以重写式(9)为

$$W(t+1) = \pi_{\eta(t)}(W(t)) = \arg\min_{W} \frac{1}{2} \| W - U(t) \|_2^2 + \frac{\lambda_2}{\eta(t)} \| W \|_{2,1} \tag{10}$$

其中: $U(t) = W(t) - \frac{1}{\eta(t)} \nabla f(W(t))$ 和 $\pi_{\eta(t)}(W(t))$ 是 $W(t)$ 在凸集 $\eta(t)$ 的欧几里德映射。由于 $W(t+1)$ 在每行的可分性,故能寻找各行封闭形式的解来得到最佳的 $W(t+1)$ [11]。

同时,为了加速式(9)中的邻近梯度方法,进一步引入一个辅助变量 $V(t+1)$:

$$V(t+1) = W(t) + \frac{\alpha(t) - 1}{\alpha(t+1)} (W(t+1) - W(t)) \tag{11}$$

其中,系数 $\alpha(t+1)$ 通常设为 $\alpha(t+1) = \frac{1 + \sqrt{1 + 4\alpha(t)^2}}{2}$。伪代码如下所示,其中算法 2 中的优化方法在定理 1 下收敛。

算法 2　优化求解式(6)的伪代码

输入: $\eta(0) = 0.01$, $\alpha(1) = 1$, $\gamma = 0.002$, ρ_1, ρ_2。
输出: W。

```
初始化 t = 1;
初始化 W(1) 作为随机对角矩阵;
重复:
    while L(W(t)) > G_{η(t-1)}(π_{η(t-1)}(W(t)), W(t))
        do 设 η(t-1) = γη(t-1);
    end
    设 η(t) = η(t-1);
    计算 W(t+1) = arg min_W G_{η(t)}(W, V(t));
    计算 α(t+1) = (1 + sqrt(1 + 4α(t)^2)) / 2;
    计算式(11);
until 式(6)收敛;
```

定理 1 [12]　假设 $\{W(t)\}$ 为算法 1 所得序列,则对于 $\forall t \geqslant 1$ 得:

$$\vartheta(W(t)) - \vartheta(W^*) \leqslant \frac{2\gamma L \| W(1) - W^* \|_F^2}{(t+1)^2} \tag{12}$$

其中: $\gamma > 0$ 是预先定义的常数; L 是式(7)中 $f(W)$ 的梯度的李普希茨常数(Lipschitz constant),此外 $W^* = \arg\min_{W} \vartheta(W)$。定理 1 表明算法 1 中的近端加速梯度法(accelerated proximal gradient method)的收敛率为 $O(\frac{1}{t^2})$,其中 t 定义迭代次数。

2　实验与结果分析

2.1　实验数据集和评价指标

本文采用六个数据集来测试算法性能,其中数据集 arcene、madelon 来源于 UCI [13]; breast cancer 来源于文献[14]; GDS1027、GDS1454 来源于 NCBI [15],数据集详情如表 1 所示。

为了评价算法的效果,本文采用分类准确率对实验结果进行衡量,分类准确率越大表明分类效果越好。

表 1　数据集信息统计

数据集	样本数	属性数	类数
arcene	100	9 920	2
breast cancer	286	22 283	2
madelon	2000	500	2
GDS1027	154	26 923	4
GDS1454	180	54 613	4
train	168	147	9

实验在 Windows 7 系统下运行,使用 MATLAB 2014a 软件进行编程、实验。实验选择六种对比算法与本文所提出的方法进行比较: NFS 方法(non feature selection),对原始数据不进行属性约简,直接使用 LIBSVM 工具箱[16]进行 SVM 分类; PCA 主成分分析方法、LDA 方法、LPP 方法、LE 方法(Laplacian eigenmaps)使用拉普拉斯特征映射对原始数据进行属性约简处理; L21 方法($l_{2,1}$-sparse),使用式(1)方法对原始数据通过稀疏处理选取新的属性子集。

分析以上算法,NFS 方法直接对原始数据集进行 SVM 分类,未对原始数据进行任何处理,相比 SG_FS 等属性约简算法,不仅数据处理量大,而且容易受到冗余数据和噪声数据的影响; PCA 考虑数据的主成分把数据从高维空间投影到低维空间; LDA 和 LPP 分别考虑了数据的全局信息和局部信息; LE 通过构建相似关系图来重构数据流形的局部结构特征[17]; L21 方法产生稀疏矩阵,由此赋值大权重给重要的属性,赋值小权重给不重要的属性。而且比较的算法中,PCA、LDA、LPP 和 LE 等算法属于子空间学习方法,L21 和本文算法属于属性选择方法。

算法复杂度方面: 本文 SG_FS 算法是稀疏学习、LDA 和 LPP 有效地组合,其时间复杂度与 LPP、LDA、LE、L21 算法一样,即 $O(n^3)$(其中 n 是样本量); 本文算法及本文的比较算法均需存储矩阵乘法的中间结果,空间复杂度为线性。

2.2　实验结果和分析

实验采用十折交叉验证法将数据分成训练集和测试集。所有算法都在同一实验环境下进行实验,每个数据集重复运行 10 次以避免实验可能产生的误差,即增加实验的稳定性。为了更好地展示各算法的性能和稳定性,本文取各算法运行 10 次的结果的平均值加或减方差来报告实验结果,统计结果如表 2 所示。另外,为了比较不同数据量纲的各算法的稳定性,本文采用变异系数(coefficient of variation)(变异系数 = (标准差/平均值) × 100%)作为评价指标,统计结果如表 3 所示。为更直观地比较各算法性能,各算法每次的实验结果对比图显示在图 1 ~ 6。

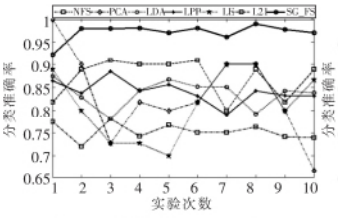

图 1　数据集 arcene 上各算法准确率对比

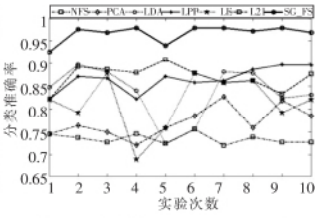

图 2　数据集 breast cancer 上各算法准确率对比

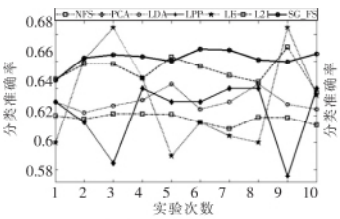

图 3　数据集 madelon 上各算法准确率对比

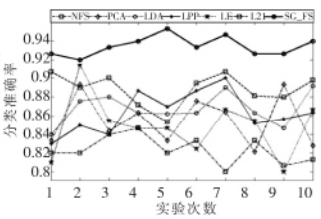

图 4　数据集 GDS1027 上各算法准确率对比

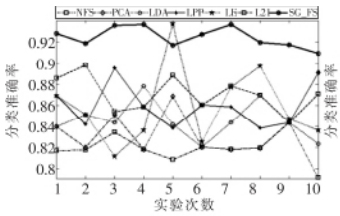

图 5 数据集 GDS1454 上各算法准确率对比

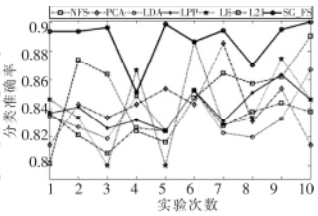

图 6 数据集 train 上各算法准确率对比

如表 2 所示，在各个数据集上 SG_FS 算法取得的准确率均为最高，且与 NFS 算法比较平均提高了 12.93%，效果最为明显，与 L21 算法比较平均提高了 5.84%。其中，在 arcene 数据集上效果最为显著，SG_FS 与其他对比算法相比，均超过 10% 左右，且与 NFS 相比提高了 21.83%。这是因为 SG_FS 算法综合了对比算法的优点，并弥补了单一方法的不足，充分考虑了数据的整体和局部信息，能有效地选择主要的类判别属性和去除噪声属性，因此能显著提高分类性能。同时，根据图 1 ~ 6 可以直观地看出，SG_FS 算法在各数据集上 10 次运行的结果的折线大部分能在其他算法的上方，表明 SG_FS 算法具有很好的分类效果。

分析表 2 的方差统计结果可看出，SG_FS 算法在各个数据集的 10 次运行结果取得的方差与各对比算法相比，均能取得最小的方差。因此，本文 SG_FS 算法比对比算法有更好的稳定性。分析表 3 的变异系数统计结果可以得出，本文 SG_FS 算法在各数据集上得到的变异系数均为最小，特别在 madelon 数据上仅为 0.1%。因此，通过分析各算法 10 次运行结果得到的方差(表 2)和变异系数(表 3)比较可知，本文 SG_FS 算法比对比算法稳定。

表 2 准确率(均值(方差)统计结果

数据集	NFS	PCA	LDA	LPP	LE	L21	SG_FS
arcene	75.33 ±4.47	83.30 ±9.20	83.75 ±8.86	84.16 ±6.21	79.28 ±15.7	87.21 ±7.94	97.16 ±3.52
breast cancer	73.48 ±1.27	77.17 ±5.06	83.64 ±11.3	86.57 ±3.92	81.55 ±11.69	87.02 ±3.03	96.68 ±1.08
madelon	61.77 ±2.15	62.25 ±5.68	63.06 ±0.50	62.25 ±1.28	63.24 ±2.64	65.31 ±0.98	66.21 ±0.42
GDS1027	82.33 ±2.19	85.51 ±7.16	86.76 ±2.86	86.35 ±4.98	84.84 ±10.3	88.82 ±2.85	93.47 ±1.07
GDS1454	81.89 ±1.93	83.91 ±6.33	84.56 ±3.19	85.96 ±4.25	85.90 ±6.90	87.08 ±2.90	92.46 ±0.89
train	83.14 ±3.59	84.15 ±4.37	83.47 ±2.63	84.02 ±1.65	84.16 ±4.40	85.01 ±3.77	89.54 ±2.38
平均	76.32	79.38	80.87	81.55	79.83	83.41	89.25

表 3 各数据集十次结果的变异系数统计结果

数据集	NFS/%	PCA/%	LDA/%	LPP/%	LE/%	L21/%	SG_FS/%
arcene	2.8	3.6	3.6	3.0	5.0	3.2	1.9
breast cancer	1.5	2.9	4.0	1.8	4.2	2.0	1.1
madelon	2.4	3.8	1.1	1.8	2.6	1.5	0.1
GDS1027	1.8	3.1	2.0	2.6	3.8	1.9	1.1
GDS1454	1.7	3.0	2.1	2.4	3.1	2.0	1.0
train	2.3	2.5	1.9	1.7	2.5	2.3	1.7

3 结束语

本文提出的属性选择算法 SG_FS 整合了线性判别分析(LDA)和局部保持投影(LPP)的思想，即考虑了数据整体和局部的信息，并有效地将子空间学习融入了现有的稀疏属性选择框架。该算法综合了稀疏学习和子空间学习的优点，并弥补基于稀疏学习的属性选择算法的不足。经实验验证，本文算法能够在分类准确率和稳定性上取得显著提高。

参考文献:

[1] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[2] Zhu Xiaofeng, Zhang Shichao, Jin Zhi, *et al*. Missing value estimation for mixed-attribute data sets [J]. IEEE Trans on Knowledge & Data Engineering, 2010, 23(1): 110-121.

[3] Zhu Xiaofeng, Suk H I, Shen Dinggang. Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 2014: 3089-3096.

[4] Zhu Xiaofeng, Huang Zi, Cheng Hong, *et al*. Sparse hashing for fast multimedia search [J]. ACM Trans on Information Systems, 2013, 31(2): 595-605.

[5] Zhu Xiaofeng, Huang Zi, Yang Yang, *et al*. Selftaught dimensionality reduction on the high-dimensional small-sized data [J]. Pattern Recognition, 2013, 46(1): 215-229.

[6] Zhu Xiaofeng, Li Xuelong, Zhang Shichao. Blockrow sparse multiview multilabel learning for image classification [J]. IEEE Trans on Cybernetics, 2015, 46(2): 450-461.

[7] Yang Yi, Shen Hengtao, Ma Zhigang, *et al*. $L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning [C]//Proc of the 22nd International Joint Conference on Artificial Intelligence. 2011: 1589-1594.

[8] Wang Hua, Nie Feiping, Huang Heng, *et al*. Identifying AD sensitive and cognition-relevant imaging bio-markers via joint classification and regression [C]//Proc of Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2011: 115-123.

[9] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Linear cross-modal hashing for efficient multimedia search [C]//Proc of the 21st ACM International Conference on Multimedia. New York: ACM Press, 2013: 143-152.

[10] Zhu Xiaofeng, Huang Zi, Cui Jiangtao, *et al*. Video-to-shot tag propagation by graph sparse group Lasso [J]. IEEE Trans on Multimedia, 2013, 15(3): 633-646.

[11] Zhu Xiaofeng, Zhang Lei, Huang Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Trans on Image Processing, 2014, 23(9): 3737-3750.

[12] Zhu Xiaofeng, Suk H I, Shen Dinggang. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis [J]. NeuroImage, 2014, 100: 91-105.

[13] UCI. UCI repository of machine learning datasets [DB/OL]. [2015-04-14]. http://archive.ics.uci.edu/ml/.

[14] Fang Xiaozhao, Xu Yong, Li Xuelong, *et al*. Locality and similarity preserving embedding for feature selection [J]. Neurocomputing, 2014, 128(5): 304-315.

[15] NCBI. NCBI dataset browser [DB/OL]. [2015-04-14]. http://www.ncbi.nlm.nih.gov/sites/GDSbrowser.

[16] Lin Chih-Jen. LIBSVM: a library for support vector machines [EB/OL]. [2015-04-14]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] Jenatton R, Gramfort A, Michel V, *et al*. Multi-scale mining of fMRI data with hierarchical structured sparsity [J]. SIAM Journal on Imaging Sciences, 2012, 5(3): 835-856.