文章编号:1672-3961(2012)04-0008-05 DOI:10.6040/j.issn.1672-3961.2012.04.002

流形学习算法在中文文本分类中的应用

王洪元,封磊,冯燕,程起才
(常州大学信息科学与工程学院,常州市过程感知与互联技术重点实验室, 江苏 常州 213164)

摘要:传统的流形学习局部线性嵌入(locally linear embedding, LLE)算法通过欧氏距离来选择邻域,如果数据集选自多个类别,这种距离度量方法无法得到正确的邻域关系。本研究提出一种改进的局部线性嵌入(modified LLE,MLLE)算法,该算法通过改进距离矩阵,使得类间的距离大、类内的距离小,从而使得邻域的选择尽量在一个类中。将MLLE算法应用到中文文本分类中,结果表明:与传统的算法比较,MLLE在分类结果可视化效果和识别率等方面都有显著提高。

关键词:流形学习;LLE算法;MLLE算法;中文文本分类
中图分类号:TN911.7 **文献标志码**:A

The manifold learning algorithm′s application in the Chinese text clustering

WANG Hong-yuan, FENG Lei, FENG Yan, CHENG Qi-cai
(Changzhou Key Laboratory for Process Perception and Interconnected Technology, School of Information Science and Engineering, Changzhou University, Changzhou 213164, China)

Abstract: According to the euclidean distance, the original LLE (locally linear embedding) algorithm chooses the neighborhood. If the data was originated from multiple classes, the correct neighborhood relationship could not be obtained. In order to solve this problem, an improved MLLE(modified LLE) was proposed. In MLLE algorithm, the distance matrix was modified, which could make the distance longger between classes and smaller within classes, and so could make the neighborhood in one class as far as possible. The test of Chinese text clustering showed that the MLLE algorithm could improve the clustering visualization and the recognition rate.

Key words: manifold learning; LLE algorithm; MLLE algorithm; Chinese text clustering

0 引言

随着Internet的迅猛发展,文本信息的数量也日益增加,文本信息的分析也随之变得重要。文本信息的分析中的一个主要技术就是文本聚类技术[1-6]。在文本挖掘领域,一般通过计算文本中词条出现的频度来构造文本-特征词矩阵进行聚类分析,而文本特征矩阵通常是高维的,这种高维的特征空间不仅会导致“维数灾”,而且让研究者难以直接理解及发现数据集的内在规律。为此,需要对文本特征矩阵进行降维处理。目前,降维方法分为线性降维和非线性降维,线性降维的方法主要有PCA[7],LDA[8-9],ICA[10],MDS[11]等;非线性降维方法有ISOMAP[12], LLE[13], LTSA[14], Hessian LLE[15],Laplacian Eigenmaps[16]等。此外文献[17]还提出了一种新的有监督的降维方法——非参数判别性局部线性嵌入(nonparametric locally linear discriminant embedding, NLLDE)将LLE和加权非参数最大间隔(weighted non-parametric maximum margin

收稿日期:2012-05-06
基金项目:国家自然科学基金资助项目(61070121)
作者简介:王洪元(1960-),男,江苏常州人,教授,博士研究生,主要研究方向为模式识别与智能系统. E-mail:hywang@cczu.edu.cn

criterion,WNMMC)进行融合。文献[18-20]提到了核主元分析法(KPCA),即通过核函数来完成非线性变换。

原始的 LLE 算法对于采样于单个流形的样本数据能够得到较好的低维嵌入,该算法首先通过 k 近邻来构造权值矩阵,然后在低维嵌入的时候也保持这种权值不变,但是这样做的结果就是可能因为近邻 k 选择不当,而造成多个子流形之间的重叠,不能得到较好的低维嵌入。综上所述,不同的近邻个数会产生不同的重构误差,因此为了得出较好的降维结果,就有必要减小重构误差,为此,本研究提出了 1 种改进局部线性嵌入算法(MLLE),通过对距离矩阵的改进来减少近邻参数对降维结果的影响。为了能更好地对文本特征矩阵进行处理,首先用 MLLE 算法对文本特征矩阵进行降维,然后再使用 LDA 算法进行二次降维,将得到的结果进行聚类可视化分析。在此基础上,本研究给出了 MLLE-LDA 的文本识别率分析并和其他几种算法进行了比较,说明了本文算法的可行性。

2　MLLE 算法

2.1 原始的 LLE 算法

LLE 认为在局部意义下数据的结构是线性的,于是任取一点,可以用其最近邻居点的线性组合来表示这一点。因而 LLE 算法的主要思想是建立原高维空间数据的邻近数据局部线性表示,通过在将为空间中尽可能保持其局部线性表示特征来实现降维。

LLE 算法的具体步骤可概括为下列 3 个步骤。

步骤 1　局部近邻选取。对于给定的数据集 $X=\{x_1,x_2,\cdots,x_N\}$,$x_i\in\mathbf{R}^{\mathrm{d}}(i=1,2,\cdots,N)$ 利用样本点间的欧式距离在高维空间中寻找每个样本点 x_i 的$k(k<N)$个近邻点。

步骤 2　计算样本点的局部重建权值,使样本点的重建误差最小。即求得以下最优问题:

$$\begin{cases}\min\ \varepsilon(\boldsymbol{W})=\sum\limits_{i=1}^{N}\|x_i-\sum\limits_{j=1}^{N}w_{ij}x_j\|_2^2,\\ \text{s.t}\sum\limits_{j=1}^{k}w_{ij}=1。\end{cases}\tag{1}$$

步骤 3　利用权值矩阵 $\boldsymbol{W}$ 寻找样本集的低维嵌入 $\boldsymbol{Y}$。通过最小化重构误差和函数 $\min\boldsymbol{\Phi}(\boldsymbol{Y})=\sum\limits_{i=1}^{N}\|y_i-\sum\limits_{j=1}^{N}w_{ij}y_i\|_2^2$ 来实现。

为了固定 $\boldsymbol{Y}$ 和避免数据集在低维坍塌到坐标原点,可简单地对 $\boldsymbol{Y}$ 加以限制:$\sum\limits_{i=1}^{N}y_i=0$,$\frac{1}{N}\sum\limits_{i=1}^{N}y_iy_i^{\mathrm{T}}=\boldsymbol{I}$。其中,$\boldsymbol{I}$ 表示 N 维单位矩阵。相应的优化问题转化为下列约束优化问题:

$$\begin{cases}\min\boldsymbol{\Phi}(\boldsymbol{Y})=\sum\limits_{i=1}^{N}\|\boldsymbol{Y}\boldsymbol{I}_i-\boldsymbol{Y}\boldsymbol{W}_i\|^2=\\ \qquad\sum\limits_{i=1}^{N}\|\boldsymbol{Y}(\boldsymbol{I}_i-\boldsymbol{W}_i)\|=\\ \qquad\min\operatorname{tr}\boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}},\\ \text{s.t.}\ \ \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{I}。\end{cases}\tag{2}$$

其中,$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})$。使用 Lagrange 乘子法,解得 $\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}=\lambda\boldsymbol{Y}^{\mathrm{T}}$,取 $\boldsymbol{M}$ 的值最小的 d 个非零特征值所对应的特征向量作为低维坐标 $\boldsymbol{Y}$。通常最小的特征值几乎为零,因此取 2 ~ $(d+1)$ 的特征值所对应的特征向量作为输出结果。

2.2　改进的局部线性嵌入(MLLE)算法

众所周知,原始的 LLE 算法对于 k 近邻的选择比较敏感,如果 k 近邻的选择的较大,会出现不同类之间的邻域重叠的现象;如果 k 近邻的选择的较小,就不能得到正确的低维嵌入,2003 年,Kouropteva O 等提出一种有监督的 LLE 算法,即 SLLE[21],SLLE 算法通过增加类间的距离而减少类内的距离来减少 k 近邻的选择对低维嵌入的影响。此后,还有相关文献[17,22-24]等都是通过改进距离矩阵的方法(用到了样本的类别信息)来选择 k 近邻,从而可以有效地提取数据的低维鉴别子流形,使得分类性能要优于非监督的维数约简方法。本研究也是通过改进距离的方法来减少近邻参数对降维结果的影响,具体的距离矩阵构造如下:

$$\boldsymbol{D}_{ij}=\begin{cases}\exp\left(-\dfrac{d^2(x_i,x_j)}{\beta}\right)\left[1-\exp\left(-\dfrac{d^2(x_i,x_j)}{\beta}\right)\right],\\ \text{if } x_i \text{ is the neighbors of } x_j \text{ and } x_i,\ x_j \text{ have the}\\ \text{same label};\\ 1-\exp\left(-\dfrac{d^2(x_i,x_j)}{\beta}\right),\ \text{if } x_i \text{ is the neighbors}\\ \text{of } x_j \text{ and } x_i,\ x_j \text{ have different label}。\end{cases}\tag{3}$$

此处 $d(x_i,x_j)$ 是样本间的欧氏距离,β 是一个调节参数。

图 1 中 L_1 代表 x_i 是 x_j 的 k 近邻,并且 x_i 与 x_j 具有相同的类标签,L_2 代表 x_i 是 x_j 的 k 近邻,同时 x_i 与 x_j 有不同的类标签。L_3 就是给出了 L_1 和 L_2 的变化范围。根据图 1 权值定义的优点可以总结如下:当距离相等时,类间点距大于类内点之间的距离。因此类间点的权值大于类内点的权值,这样对于分类是有效的;样本 x_i 是 x_j 的 k 近邻,如果二者有相同的标签,则二者之间的权重使其分布更近。

然而,如果其标签不同,这个权值将使其相距更远。

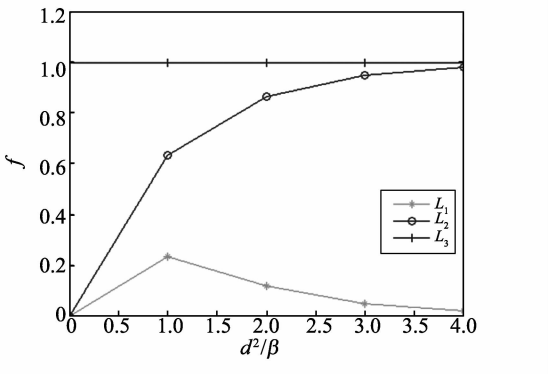

图 1 距离关系图

Fig. 1 Distance relationship chart

3 MLLE-LDA 算法在中文文本分类中的应用

3.1 实验和结果分析

3.1.1 实验样本集

为了检验算法的可行性和有效性,采用谭松波等人建立的中文文本分类语料库 TanCorp 在 matlab7 上进行仿真。该语料数据采用词频矩阵的方式给出,词表中特征词个数为 72 641,分为人才,体育,卫生,地域,娱乐,房产,教育,汽车共 12 大类,从中抽取库中 4 个类别,分别为人才,房产,教育和汽车,每个类别再抽取前 100 个文档数据进行实验,形成一个 400 × 72 641 维的文本 - 特征词矩阵。

3.1.2 实验结果

本实验将提出的 MLLE-LDA 算法与 PCA 算法、ISOMAP 算法和 LLE 算法的实验结果进行比较,本研究经过反复实验认为选取 k 邻域参数为 100 比较合适(如果邻域参数大于 100,就会出现短路边现象,实验结果会重叠;如果小于 100,则不能观察到数据的内部结构)。实验图 2 中的(a)、(b)、(c)、(d)代表的是各种算法降维后的 2 维可视化效果图。

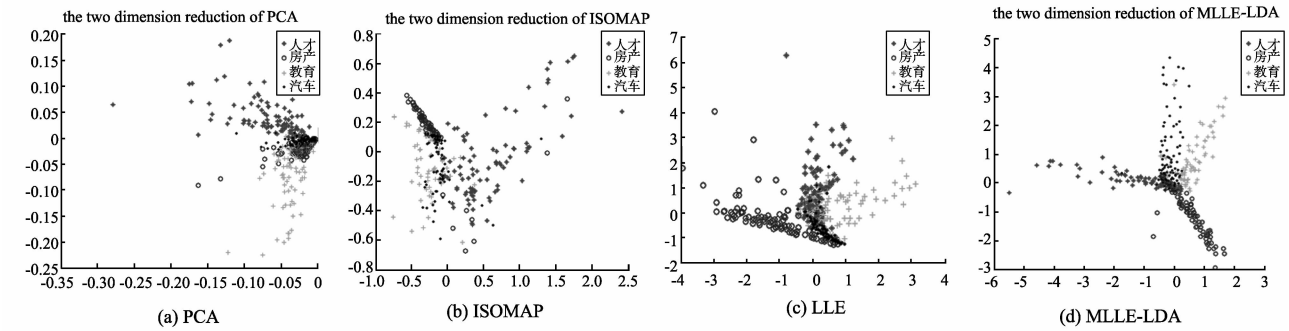

图 2 各种算法降维后的可视化效果图

Fig. 2 The visualization figures for several algorithms

由图 2(a)、(b)、(c)可以看出,PCA 算法、ISOMAP 算法和 LLE 算法不能将 4 类数据完全展开,出现了类内重叠现象,不能保持数据的内部结构;而从图 1(d)的实验结果可以看出本研究提出的 MLLE-LDA 算法成功地将四类数据展开,并得到了很好的可视化效果。

4 MLLE-LDA 文档识别

4.1 各个方法识别文档的比较研究

对本研究中提取的数据进行简单的归一化,得到一个分布于 0 和 1 之间的 400 × 72 641 维的文本 - 特征词矩阵。首先使用 LLE 进行降维处理得到一个 400 × 10 的特征矩阵,然后通过十字交叉验证法将得到的特征矩阵选取 90% 作为测试集,而 10% 作为训练集,再使用 LDA 将相应的特征矩阵将至 2 维,最后使用 KNN 算法进行分类,得到相应的类标签,并与其他分类算法进行比较,比较结果见表 1。

表 1 各种方法分类的正确率

Table 1 The classification accuracy for several algorithms

方法	次数										均值
	1	2	3	4	5	6	7	8	9	10	
贝叶斯分类	0.515 0	0.522 5	0.520 0	0.520 0	0.497 5	0.507 5	0.527 5	0.512 5	0.517 5	0.525 0	0.516 5
LLE	0.300 0	0.400 0	0.275 0	0.375 0	0.275 0	0.200 0	0.375 0	0.375 0	0.325 0	0.375 0	0.327 5
ISOMAP	0.763 2	0.736 8	0.684 2	0.763 2	0.648 6	0.605 3	0.717 9	0.657 9	0.589 7	0.584 6	0.675 1
PCA	0.700 0	0.750 0	0.650 0	0.800 0	0.725 0	0.825 0	0.625 0	0.650 0	0.750 0	0.625 0	0.710 0
MLLE + LDA	0.850 0	0.675 0	0.750 0	0.850 0	0.800 0	0.825 0	0.850 0	0.800 0	0.875 0	0.875 0	0.815 0

表 1 是通过 10 次十字交叉得到的结果,图 3 是根据表 1 中的数据得到识别率分布图,从实验结果可以明显看出 LLE 算法的识别效果最低,贝叶斯方法的识别效果其次,ISOMAP 和 PCA 算法又优于前两种算法,而本文提出 MLLE-LDA 算法的识别率是最高的。

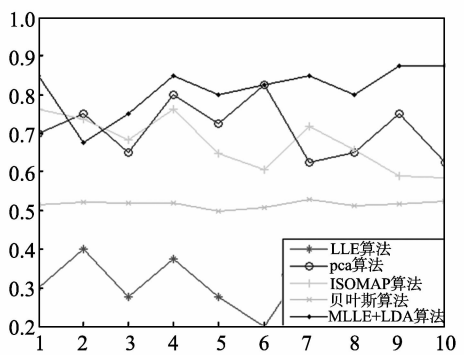

图 3　各种方法的识别率分布图

Fig. 3　The recognition rate distribution for the studied algorithms

4.2　评估标准

F 度量[25]是一种参照信息检索的评测方法,将每个聚类结果看作是查询的结果,这样对于最终的某一个聚类类别 r 和原来的预定义类别 i,得出:

准确率: $\text{precision}(i,r) = n_{ir}/n_r$,　　(5)

召回率: $\text{recall}(i,r) = n_{ir}/n_i$。　　(6)

这里,n_{ir} 是聚类 r 中包含类别 i 中的文本的个数,n_r 是聚类类别 r 中实际对象的数目,n_i 是原来预定义类别 i 应有文本数。则聚类 r 和类别 i 之间的 f 值计算如式(7)。

$$f(i,r) = \frac{2\text{recall}(i,r) * \text{precision}(i,r)}{\text{recall}(i,r) + \text{precision}(i,r)}。\quad (7)$$

给出了分类的各项评估标准:准确率、召回率和 F_1 估计值,因为提取文本 - 特征词矩阵的来自人才、房产、教育、汽车这 4 类,所以对这 4 类的分类效果进行了统计。从表 2 中可以看出,在这几种方法中,在与贝叶斯算法、PCA 算法、LLE 算法、ISOMAP 算法取得相同精度的情况下,MLLE-LDA 的召回率(查全率)是最高的,而通过准确率和召回率折合的 F 度量也是本研究提出的 MLLE-LDA 算法较优。

表 2　F 度量

Table 2　F measure

算法类别	准确率					召回率					F_1 估计值				
	贝叶斯	PCA	LLE	ISOMAP	MLLE + LDA	贝叶斯	PCA	LLE	ISOMAP	MLLE + LDA	贝叶斯	PCA	LLE	ISOMAP	MLLE + LDA
人才	0.140	0.319	0.283	0.308	0.243	0.296	0.900	0.370	0.800	0.800	0	0.470	0.311	0.444	0.372
房产	0.173	0.218	0.209	0.297	0.267	0.359	0.620	0.270	0.770	0.880	0	0.320	0	0.429	0.410
教育	0.284	0.226	0.192	0.170	0.276	0.588	0.640	0.230	0.450	0.910	0.380	0.330	0.206	0.247	0.424
汽车	0.406	0.237	0.316	0.225	0.214	0.823	0.670	0.390	0.590	0.710	0.540	0.350	0.341	0.326	0.329

5　结论

文本数据集内部可能包括多个类别,本研究根据文本数据集的这种特性以及 LLE 算法的相关优点,提出了 MLLE-LDA 算法对文本分类进行处理。从文本分类的实验结果来看,比较其他降维算法,本研究提出的算法可从一定程度上产生较好的可视化效果,与此同时,其相应的文档分类准确率精度也更高。虽然本研究从一定程度上减小了邻域参数对降维效果的影响,但是并没有彻底解决这个问题,所以消除邻域参数对降维效果的影响是以后仍要继续研究的内容。

参考文献:

[1] 马帅, 王腾蛟, 唐世渭, 等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报, 2003, 14(6):1089-1095.
MA Shuai, WANG Tengjiao, TANG Shiwei, et al. A fast clustering algorithm based on reference and density [J]. Journal of Software, 2003, 14(6):1089-1095.

[2] 王玲, 薄列峰, 焦李成. 密度敏感的半监督谱聚类[J]. 软件学报, 2007, 18(10):2412-2422.
WANG Ling, BO Liefeng, JIAO Licheng. Density-sensitive semi-supervised spectral clustering [J]. Journal of Software, 2007, 18(10):2412-2422.

[3] 彭京, 杨冬青, 唐世渭, 等. 一种基于语义内积空间模型的文本聚类算法[J]. 计算机学报, 2007, 30(8):1354-1363.
PENG Jing, YANG Dongqing, TANG Shiwei, et al. A novel text clustering algorithm based on inner product space model of semantic[J]. Chinese Journal of Computers, 2007, 30(8):1354-1363.

[4] 孙学刚, 陈群秀, 马亮. 基于主题的 Web 文档聚类研究[J]. 中文信息学报, 2003, 17(3):21-26.
SUN Xuegang, CHEN Qunxiu, MA Liang. Study on topic-based web clustering[J]. Journal of Chinese Information Processing, 2003, 17(3):21-26.

[5] 吴斌, 傅伟鹏, 史忠植, 等. 一种基于群体智能的 web 文档聚类算法[J]. 计算机研究与发展, 2002, 39(11):

1429-1435.
WU Bin, FU Weipeng, SHI Zhongzhi, et al. A clustering algorithm based on swarm intelligence for web document[J]. Journal of Computer Research and Development, 2002, 39(11):1429-1435.

[6] 寇苏玲, 蔡庆生. 中文文本分类中的特征选择研究[J]. 计算机仿真, 2007, 24(3):289-291.
KOU Suling, CAI Qingsheng. Research on feature-selection in Chinese text classification[J]. Computer Simulation, 2007, 24(3):289-291.

[7] YEUNG K Y, RUZZO W L. Principal component analysis for clustering gene expression data[J]. Bioinformatics, 2001, 17(9):763-774.

[8] YU H, YANG J. A direct LDA algorithm for high-dimensional data with application to face recognition[J]. Pattern Recognition, 2001, 34:2067-2070.

[9] 张玉华, 王欣. 基于线性判别分析的加权零空间算法及在人脸识别中的应用[J]. 山东大学学报:工学版, 2009, 39(6):31-34.
ZHANG Yuhua, WANG Xin. A LDA-based weighted null space algorithm in face recognition[J]. Journal of Shandong University: Engineering Science, 2009, 39(6):31-34.

[10] HYVARINEN A. Fast and robust fixed-point algorithms for independent component analysis[J]. Neural Networks, 1999, 10(3):626-634.

[11] CHEN L, BUJA A. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis[J]. Journal of the American Statistical Association, 2009, 104(485):209-219.

[12] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500):2319-2323.

[13] ROWEIS S T. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500):2323-2326.

[14] ZHANG Z, ZHA H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment[J]. Journal of Shanghai University: English Edition, 2004, 8(4):406-424.

[15] DONOHO D L, GRIMES C. Hessian eigenrnaps:locally linear embedding techniques for high-dimensional data[J]. National Academy of Sciences of the United States of America, 2003, 100(10):5591-5596.

[16] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15:1373-1396.

[17] 王熙照, 白丽杰, 花强, 等. 基于非参数判别性分析的局部线性嵌入算法研究[J]. 山东大学学报:工学版, 2011, 41(4):1-6.
WANG Xizhao, BAI Lijie, HUA Qiang, et al. Locally linear discriminant embedding with nonparametric method[J]. Journal of Shandong University:Engineering Science, 2011, 4(41):1-6.

[18] 颜子夜,陆耀,李建武, 等. 一种基于核主成分分析的图像超分辨率算法[J]. 山东大学学报:工学版, 2005, 35(3):103-106.
YAN Ziye, LU Yao, LI Jianwu, et al. Kernel principal components analysis based super resolution method[J]. Journal of Shandong University: Engineering Science, 2005, 35(3):103-106.

[19] 崔燕,范丽亚. 高维数据正定核与不定核的 KPCA 变换阵比较[J]. 山东大学学报:工学版, 2011, 41(1):17-23.
CUI Yan, FAN Liya. Comparison of KPCA transformation matrices with definite and indefinite kernels for high-dimensional data[J]. Journal of Shandong University: Engineering Science, 2011, 41(1):17-23.

[20] 邓晓刚, 田学民. 一种基于 KPCA 的非线性故障诊断方法[J]. 山东大学学报:工学版, 2005, 35(3):103-106.
DENG Xiaogang, TIAN Xuemin. Nonlinear process fault diagnosis method using kernel principal component analysis[J]. Journal of Shandong University: Engineering Science, 2005, 35(3):103-106.

[21] KOUROPTEVA O, OKUN O, PIETIKÄINEN M. Supervised locally linear embedding algorithm for pattern recognition[J]. Pattern Recognition and Image Analysis, 2003, 2652:386-394.

[22] 王和勇, 郑杰, 姚正安, 等. 基于聚类和改进距离的 LLE 方法在数据降维中的应用[J]. 计算机研究与发展, 2006, 43(8):1485-1490.
WANG Heyong, ZHENG Jie, YAO Zheng′an, et al. Application of dimension reduction on using improved LLE based on clustering[J]. Journal of Computer Research and Development, 2006, 43(8):1485-1490.

[23] 陆建新, 李宏宇, 沈一帆, 等. 一种改进的局部线性嵌套方法[J]. 计算机应用与软件, 2008, 25(10):9-10.
LU Jianxin, LI Hongyu, SHEN Yifan, et al. An improved algorithm on locally linear embedding[J]. Computer Applications and Software, 2008, 25(10):9-10.

[24] 温金环, 田铮, 林伟, 等. 基于监督局部线性嵌入特征提取的高光谱图像分类[J]. 计算机应用, 2011, 31(3):715-717.
WEN Jinhuan, TIAN Zheng, LIN Wei, et al. Feature extraction based on supervised locally linear embedding for classification of hyperspectral images[J]. Journal of Computer Applications, 2011, 31(3):715-717.

[25] 索红光, 王玉伟. 一种用于文本聚类的改进 k-means 算法[J]. 山东大学学报:理学版, 2008, 43(1):1-5.
SUO Hongguang, WANG Yuwei. An improved k-means algorithm for document clustering[J]. Journal of Shandong University: Natural Science, 2008, 43(1):1-5.

（编辑:胡春霞）