中图法分类号: TP391.41 **文献标志码**: A **文章编号**: 1006-8961(2011)09-1615-10
论文索引信息:尹学松,胡恩良. 半监督局部维数约减[J]. 中国图象图形学报,2011,16(9):1615-1624

半监督局部维数约减

尹学松[1),2)],胡恩良[1)]

[1)](南京航空航天大学信息科学与技术学院, 南京 210016) [2)](浙江广播电视大学信息与工程学院, 杭州 310030)

摘 要:在挖掘和分析高维数据任务中,有时只能获得有限的成对约束信息(must-link 约束和 cannot-link 约束),由于缺乏数据类标号信息,监督维数约减方法常常不能得到满意的结果。在这种情况下,使用大量的无标号样本可以提高算法的性能。文中借助于成对约束信息和大量无标号样本,提出半监督局部维数约减方法(SLDR)。SLDR 集成数据的局部信息和成对约束寻找一个最优投影,当数据被投影到低维空间时,不仅 cannot-link 约束中样本点对之间距离更远、must-link 约束中样本点对之间距离更近,数据的内在几何信息还被保持。而且 SLDR 能推广为非线性方法,使之能够适应非线性数据的维数约减。在各种数据集上的实验结果充分验证了所提出算法的有效性。

关键词: 成对约束;局部信息;维数约减;判别分析算法

Semi-supervised locality dimensionality reduction

Yin Xuesong[1),2)], Hu Enliang[1)]

[1)](*College of Information Science and Technology*, *Nanjing University of Aeronautics & Astronautics*, *Nanjing* 210016 *China*)

[2)](*School of Information and Engineering*, *Zhejiang Radio & TV University*, *Hangzhou* 310030 *China*)

Abstract: In mining and analyzing high-dimensional data task, when only a small number of pairwise constraints including must-link and cannot-link are available, supervised dimensionality reduction methods tend to perform poorly due to the lack of data labels. In such cases, unlabeled samples could be useful in improving the performance. In this paper, we propose a novel semi-supervised locality dimensionality reduction algorithm (SLDR) in terms of pairwise constraints and abundant unlabeled samples. Specifically, SLDR can effectively use local information of the data and pairwise constraints to find a projection. After the data is projected onto a low-dimensional space, instances involved by cannot-link constraints are far apart, while instances involved by must-link constraints are close to each other. Moreover, the intrinsic geometric information of the data is preserved. In addition, SLDR can be extended to nonlinear dimensionality reduction scenarios by the kernel trick, which is applied to reduce the dimensions of highly nonlinear data.

Keywords: pairwise constraint; locality information; dimensionality reduction; discriminant analysis

0 引 言

在高维空间里,由于维数灾难,难以对数据进行分类或者聚类等操作。只有当高维数据被投影到一个低维空间后,各种后续工作才能顺利进行。维数约减的目的是将高维数据投影到低维空间的同时,尽可能多地保持数据的内在信息。监督维数约减算法[1-2]和无监督维数约减算法[3-6]是维数约减算法的两种重要形式。通常,当获得大量的类标号信息

收稿日期:2009-11-26;**修回日期**:2010-05-13

基金项目:浙江省自然科学基金项目(Y1100349);浙江省教育厅项目(Z201017701);浙江省高等学校优秀青年教师项目(2008)。

第一作者简介:尹学松(1975—),男。计算机应用技术专业博士研究生,主要研究方向为数据挖掘和机器学习。E-mail: yinxs@nuaa.edu.cn。

时,监督方法优于无监督方法。然而,获得大量的类标号信息不仅费时费力,有时甚至要付出相当大的代价,如会谈中说话人语音的分割与识别[7],GPS 数据中的道路检测[8]和电影片段中不同男演员或女演员的分组[9]等问题,这就限制了监督维数约减方法的应用。半监督维数约减方法是介于监督方法和无监督方法之间,它使用一部分辅助信息来减少数据维数。一方面,不同于监督维数约减方法需要大量的类标号,它只需要一小部分先验信息来引导维数约减;另一方面,它能提高无监督方法的质量。因此,半监督维数约减方法已经得到众多研究人员的关注[10-15],成为机器学习和数据挖掘领域中的重要研究内容之一。

半监督维数约减方法使用的辅助信息有多种形式,如类标号信息、成对约束等。与得到数据的类标号相比,得到成对约束更容易[7,10,12,16-18]。因为对一个专家或者用户来说,指出两个样本是否属于相同类或者不同类比指出样本属于哪一类要相对容易。而且由数据类标号能得到成对约束,但不能由成对约束来得到数据的类标号。成对约束有 must-link 约束和 cannot-link 约束两种形式,它们定义如下[8-9,12,18]。

一个 must-link 约束规定:两个样本属于相同类;

一个 cannot-link 约束规定:两个样本属于不同类。

现有的基于成对约束的半监督维数约减算法,或者只使用成对约束来减少数据维数[10,15,17],忽略了大量的无标号样本;或者忽略了数据的局部结构[12]。为了解决上述问题,文中提出一个新颖的半监督局部维数约减方法(SLDR)。详细地说,新算法集成成对约束和数据的局部信息到一个联合框架中,因此,当数据被投影到低维空间时,不仅 cannot-link 约束中样本点对之间距离更远、must-link 约束中样本点对之间距离更近,数据的内在几何信息还被保持。事实上,在缺乏大量类标号信息的情况下,对高维数据进行维数约减时,数据的局部信息往往要比全局信息更重要[19-20]。

1 相关工作

给定一个由 n 个样本组成的数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\in\mathbf{R}^d$,线性维数约减方法需要找到一个变换矩阵 $\boldsymbol{A}=[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_r]\in\mathbf{R}^{d\times r}$,将 n 个样本投影到低维空间里

$$y_i=\boldsymbol{A}^{\mathrm{T}}x_i\in\mathbf{R}^r\quad r<d \tag{1}$$

式中,$\boldsymbol{A}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{A}$ 的转置。

大多数半监督维数约减方法可以概括为优化下面的问题

$$A^*=\arg\max\frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\bar{\boldsymbol{S}}\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\underline{\boldsymbol{S}}\boldsymbol{A})} \tag{2}$$

式中,$\mathrm{tr}(\boldsymbol{B})$ 是求解矩阵 $\boldsymbol{B}$ 的迹。另外,如果 $\bar{\boldsymbol{S}}$ 和 $\underline{\boldsymbol{S}}$ 分别是类间散布矩阵和类内散布矩阵,则它们的定义为

$$\bar{\boldsymbol{S}}=\sum_{c=1}^{C}n_c(\boldsymbol{u}_c-\boldsymbol{u})(\boldsymbol{u}_c-\boldsymbol{u})^{\mathrm{T}}$$

$$\underline{\boldsymbol{S}}=\sum_{c=1}^{C}\sum_{x_i\in l_c}(\boldsymbol{x}_i-\boldsymbol{u}_c)(\boldsymbol{x}_i-\boldsymbol{u}_c)^{\mathrm{T}}$$

求解上述最优的投影矩阵,等价于求解下面的广义特征值问题

$$\bar{\boldsymbol{S}}\boldsymbol{a}=\lambda\underline{\boldsymbol{S}}a \tag{3}$$

式(3)前 r 个最大特征值对应的特征向量,构成所求的投影矩阵 $\boldsymbol{A}=[a_1,\cdots,a_r]$。

1.1 半监督判别分析

通过使用一部分有标号的数据,Cai 等人[11]将线性判别分析算法引入半监督领域中,提出一种半监督判别分析(SDA)算法。SDA 优化的目标函数是

$$A^*=\arg\max\frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{S}_t+\lambda\boldsymbol{S}_l)\boldsymbol{A})} \tag{4}$$

式中,$\boldsymbol{S}_l$ 定义为

$$\boldsymbol{S}_l=\frac{1}{2}\sum_{i,j=1}^{n}S_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}} \tag{5}$$

$\boldsymbol{S}_{ij}$ 度量两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似性,$\boldsymbol{S}_b$ 是给定一部分有标号样本的类间散布矩阵。Cai 等人近似地使用 $\boldsymbol{S}_t=\boldsymbol{S}_b+\boldsymbol{S}_w$ 代替 $\boldsymbol{S}_w$。

1.2 半监督维数约减框架

Song 等人[14]利用部分有类标号样本来引导维数约减过程,提出一个半监督维数约减框架。在这个框架里他们分别优化两个目标函数。他们优化的第一个目标函数是

$$A^*=\arg\max\frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{S}_w+\lambda_1\boldsymbol{S}_l+\lambda_2\boldsymbol{I})\boldsymbol{A})} \tag{6}$$

不难发现,式(6)实质上就是在 SDA 方法基础上添加了一项 Tikhonov 正则化。Song 等人优化的第 2 个目标函数是

$$A^* = \arg\max \frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{S}_b - \lambda_1 \boldsymbol{S}_w - \lambda_2 \boldsymbol{S}_l)\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})} \quad (7)$$

显然,式(7)是半监督化的最大间隔标准(MMC)。

1.3　半监督马氏度量学习(LMDM)

不同于上面两个方法,使用部分有类标号的数据来求解投影方向,Xiang 等人[17]利用对给定的 must-link 约束和 cannot-link 约束,学习最优的投影,并将学到的投影用于数据的分类和聚类中。为了求解变换矩阵 $\boldsymbol{A}$,他们的方法是首先计算 must-link 约束中所有点对之间距离的和,即

$$D_w = \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in \boldsymbol{M}} (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_j) = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{S}}_w\boldsymbol{A}) \quad (8)$$

式中

$$\hat{\boldsymbol{S}}_w = \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in \boldsymbol{M}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \quad (9)$$

相应地,cannot-link 约束中所有点对之间距离的和表示为

$$D_b = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{S}}_b\boldsymbol{A}) \quad (10)$$

式中

$$\hat{\boldsymbol{S}}_b = \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in \boldsymbol{C}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \quad (11)$$

$\boldsymbol{M}$ 和 $\boldsymbol{C}$ 分别表示 must-link 约束和 cannot-link 约束集合。Xiang 等人的方法通过优化下面的目标函数来求解最优的投影矩阵

$$\boldsymbol{A}^* = \arg\max_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}} \frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{S}}_b\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{S}}_w\boldsymbol{A})} \quad (12)$$

1.4　基于球形 K 均值的特征投影

相似于 Xiang 等人的方法,Tang 等人[10]使用成对约束来减少数据的维数,提出基于球形 K 均值的特征投影(SCREEN)方法。SCREEN 优化下面的目标函数

$$A^* = \arg\max \frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{S}}_b\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})} \quad (13)$$

式中

$$\hat{\boldsymbol{S}}_b = \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in \boldsymbol{C}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \quad (14)$$

在求出投影矩阵 $\boldsymbol{A}$ 后,SCREEN 使用基于球形 K-均值算法对所有样本聚类。该方法仅仅使用成对约束来寻找高维数据的低维流形。显然,SCREEN 和 LMDM 的共同缺点是不仅忽略大量无标号样本对维数约减的贡献,也没有考虑到数据的局部几何结构。因此,在他们方法中得到的解并不是最优的投影方向。

此外,相似于 Cai 等人的方法,Sugiyama 等人[13]利用部分有类标号样本,提出一个半监督线性判别方法。Zhang 等人[12]将成对约束引入 PCA 中,提出一种半监督相关成分分析算法。Liu 等人[19]提出一种基于图的半监督线性维数约减方法。该方法使用部分有类标号样本分别构造类内图和类间图来得到最优投影方向。类似于 Liu 等人[19]的思想,Zhao 等人[21]使用部分有类标号样本提出一种半监督局部敏感特征选择算法。

2　半监督局部维数约减

数据的局部几何结构在维数约减算法中扮演着至关重要的作用[20-23]。最近研究表明,在缺乏大量类标号信息的情况下,降低数据维数时,数据的局部结构往往要比全局结构重要[19-20]。因此,本文提出一种新的半监督局部维数约减算法(SLDR)。新方法使用成对约束和数据的局部信息来寻找最优的嵌入空间。

2.1　数据的局部几何结构

半监督学习算法的关键之处是一致性先验假设[19-21]。对分类或者聚类来说,数据的局部近邻关系意味着近邻样本有相同的类标号或者聚类标号;对维数约减来说,数据的局部近邻关系意味着近邻样本在投影低维空间时有相似的嵌入。给定一个样本集 $\boldsymbol{X}$,构建一个 k 近邻的邻接图 $\boldsymbol{G}$ 来建模近邻样本之间的关系。具体地说,如果图中两个顶点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 互为近邻,那么它们之间就存在一条边,相应的权值矩阵为 $\boldsymbol{P}$,其定义如下

$$P_{ij} = \begin{cases} \mathrm{e}^{\frac{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}} & \boldsymbol{x}_i \in \boldsymbol{N}_k(\boldsymbol{x}_j) \text{ 或 } \boldsymbol{x}_j \in \boldsymbol{N}_k(\boldsymbol{x}_i) \\ 0 & \text{其他} \end{cases} \quad (15)$$

式中,$\boldsymbol{N}_k(\boldsymbol{x}_i)$ 表示样本 $\boldsymbol{x}_i$ 的 k 近邻集合。根据上述定义,如果两个样本之间存在一条边,那么这两个样本属于近邻样本。因此,高维空间中的两个近邻样本被投影到低维流形时,自然地期望这两个仍保持近邻。为了达到这个目的,最小化下列目标函数

$$J_L(a) = \sum_{i,j} (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 P_{ij} \quad (16)$$

最小化上述目标函数,在得到低维流形的同时,保持了数据的局部几何结构[5,22]。

2.2　半监督局部维数约减算法

给定 must-link 约束和 cannot-link 约束,本文的

目标是学习一个正交投影使得在低维空间中 must-link 约束里点对之间的距离尽可能小，而 cannot-link 约束里点对之间的距离尽可能大。为此，最小化下列目标函数

$$J_{MC}(\boldsymbol{a}) = \frac{1}{2n_{\boldsymbol{M}(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{M}}}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 - \frac{1}{2n_{\boldsymbol{C}(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{C}}}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2$$
$$\text{s.t.} \quad \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} = 1 \tag{17}$$

式中，n_M 和 n_C 分别表示 must-link 约束和 cannot-link 约束中点对的个数。

式(17)得到的投影矩阵仅仅考虑到成对约束信息，而忽略了大量的无标号样本对求解投影方向的贡献。因此，得到的 $\boldsymbol{a}$ 并不是最优的投影方向。为了兼顾无标号样本以及样本的局部信息对求解最优投影方向的贡献，SLDR 集成了式(16)和(17)，最小化下面的目标函数

$$J(\boldsymbol{a}) = \frac{1}{2n}\sum_{i,j}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2P_{ij} - \frac{1}{2n}\sum_{i,j}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 + \frac{1}{2n_{M(x_i,x_j)\in M}}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 - \frac{1}{2n_{C(x_i,x_j)\in C}}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 \tag{18}$$

式(18)的第 1 项描述了在低维空间中无标号样本的近邻关系，其实质等价于局部保持投影；第 2 项表达的是变换空间中所有样本之间的平均距离，其实质等价于主成分分析标准。不难发现，式(18)是式(17)的扩展。这种扩展的优点表现在：一方面式(18)是利用无标号样本以及样本的局部信息来求解投影方向，提高算法性能；另一方面，在成对约束较少时，相对于式(17)而言，式(18)能得到更稳定的解。

为了更加直观化 SLDR 的目标函数，式(18)能改写成下列形式

$$J(\boldsymbol{a}) = \frac{1}{2}\sum_{i,j}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2S_{ij} \tag{19}$$

式中

$$S_{ij} = \begin{cases} \dfrac{P_{ij}-1}{n} + \dfrac{1}{n_M} & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{M} \\ \dfrac{P_{ij}-1}{n} - \dfrac{1}{n_C} & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{C} \\ \dfrac{P_{ij}-1}{n} & \text{其他} \end{cases} \tag{20}$$

既然权值矩阵 $\boldsymbol{P}$ 是对称阵，从式(20)可得矩阵 $\boldsymbol{S}$ 也是对称阵，即 $S_{ij} = S_{ji}$。进一步简化式(19)为

$$J(\boldsymbol{a}) = \frac{1}{2}\sum_{i,j}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2S_{ij} = \frac{1}{2}\sum_{i,j}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{a} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a} - 2\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a})S_{ij} = \frac{1}{2}(\sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{a} + \sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_jS_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a} - 2\sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a}) = \frac{1}{2}(2\sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{a} + -2\sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a}) = \sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{a} - \sum_{i,j}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{a} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}$$

式中，矩阵 $\boldsymbol{D}$ 是对角阵，且 $D_{ii} = \sum_j S_{ij}$，$\boldsymbol{L}$ 被称为拉普拉斯矩阵，其值为 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$。因此，SLDR 的目标函数被简写为

$$\min \quad J(\boldsymbol{a}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}$$
$$\text{s.t.} \quad \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} = 1 \tag{21}$$

显然，最小化这个目标函数，最优的投影向量是由求解下面式子的广义特征值问题得到

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} = \lambda\boldsymbol{a} \tag{22}$$

式(22)前 r 个最小特征值对应的特征向量，构成所求的投影矩阵 $\boldsymbol{A} = [a_1, \cdots, a_r]$。

基于上述分析，半监督局部维数约减方法(SLDR)如下：

输入　样本集 $\boldsymbol{X}$，must-link 和 cannot-link 约束集合；

输出　一个投影矩阵 $\boldsymbol{A} \in \mathbf{R}^{d\times r}$；

1)构建一个无向图 $\boldsymbol{G}$，并使用式(15)计算权值矩阵 $\boldsymbol{P}$；

2)根据给定的成对约束集合 $\boldsymbol{M}$、$\boldsymbol{C}$，以及 $\boldsymbol{P}$，使用式(20)计算矩阵 $\boldsymbol{S}$；

3)根据 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$，计算拉普拉斯矩阵 $\boldsymbol{L}$；

4)计算式(22)的广义特征值，输出 $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_r]$。

与其他几个算法[10-14,17,19-20,22-23]相似，本文所提出的 SLDR 主要计算时间用在求解广义特征值问题上。由于求解广义特征值问题的时间复杂度为 $O(n^3)$[24]，因此，本文算法的时间复杂度为 $O(n^3)$。在后面的实验中不难发现，尽管 SLDR 与相关算法[1,10,15]有相同的时间复杂度，但 SLDR 的

性能比它们的性能要好。

根据上述 SLDR 算法的理论分析,得到下面的观察:

1)SLDR 算法比 Xiang 等人[15]提出的 LMDM 算法和 Tang 等人[1]提出的 SCREEN 的算法有更一般的刻画。与 LMDM 和 SCREEN 仅使用成对约束来引导维数约减相比,SLDR 不仅继承了它们的优点,还使用了大量的无标号样本来寻找投影方向,以至于 SLDR 能得到最优的投影矩阵。另外,当不使用无标号样本时,SLDR 与 LMDM 和 SCREEN 有相似的刻画,即

$$S_{ij} = \begin{cases} \dfrac{1}{n_M} & (x_i, x_j) \in \boldsymbol{M} \\ -\dfrac{1}{n_C} & (x_i, x_j) \in \boldsymbol{C} \end{cases} \tag{23}$$

因此,LMDM 算法与 SCREEN 算法实质上是 SLDR 算法的一种特殊形式。

2)SLDR 算法有更稳定的解。在成对约束较少时,LMDM 与 SCREEN 仅使用成对约束来求解投影向量,容易造成过拟合。由于 SLDR 在求解投影方向过程中使用了大量的无标号样本,因此,新算法能够克服过拟合现象,能够得到更稳定的解。

3)由于 SLDR 在求解投影方向过程中使用了数据的局部信息,因此,与 Zhang 等人[10]提出的半监督维数约减算法(SSDR)相比,新算法对野值点不敏感。而且新算法简单、易执行,能容易地推广到 Hilbert 空间。

4)虽然 SLDR 在形式上相似于拉普拉斯谱图嵌入算法[6,12],但它们之间有着本质的区别。首先,拉普拉斯谱图嵌入算法是仅使用无标号样本进行维数约减的无监督算法,SLDR 是使用无标号样本和成对约束信息进行维数约减的半监督算法;其次,它们的权值矩阵 $\boldsymbol{S}$ 有不同的构造形式;最后,在它们的目标函数中,约束项也不相同。

2.3　核化的半监督局部维数约减算法

SLDR 是简单、易执行的线性算法,它能容易地推广到非线性维数约减算法中,降低高维非线性数据维数。

假设存在一个非线性映射 $\boldsymbol{\phi}$,使输入空间被映射到 Hilbert 空间,即 $\boldsymbol{\phi}:\boldsymbol{R}^n \to \boldsymbol{H}$。相应地,在 Hilbert 空间,数据矩阵表示成 $\boldsymbol{\phi}(\boldsymbol{X}) = [\boldsymbol{\phi}(x_1), \cdots, \boldsymbol{\phi}(x_n)]$,式(22)中求解特征向量问题有如下表达

$$\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{v} = \lambda \boldsymbol{v} \tag{24}$$

为了将 SLDR 推广到非线性情况,需要用每两个数据点内积的形式来刻画它。因此,通过计算 Hilbert 空间中每两个数据点的内积,核函数可以隐式指定非线性映射,即

$$K(x_i, x_j) = (\boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j)) = \boldsymbol{\phi}(x_i)^{\mathrm{T}}\boldsymbol{\phi}(x_j)$$

根据 Representer 定理,式(24)中的特征向量 $\boldsymbol{v}$ 可以表达成 $\boldsymbol{\phi}(x_1), \cdots, \boldsymbol{\phi}(x_n)$ 的线性组合,即存在一组系数 α_i,使得向量 $\boldsymbol{v}$ 描述为

$$\boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(x_i) = \boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{\alpha} \tag{25}$$

式中,$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \alpha_n]^{\mathrm{T}} \in \mathbf{R}^n$。

将式(25)代入式(24)中,并化简,可以得到下面的特征向量问题

$$\boldsymbol{KLK\alpha} = \lambda \boldsymbol{K\alpha} \tag{26}$$

式中,$\boldsymbol{K}$ 是核矩阵,$K_{ij} = (x_i, x_j)$。让列向量 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_r$ 是式(26) r 个最小特征值对应的特征向量,构成所求 $\boldsymbol{n \times r}$ 的投影矩阵 $\boldsymbol{\psi} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_r]$。在 Hilbert 空间中一个测试样本 $\phi(x)$,SLDR 将它投影到 r 维空间中的表达式为

$$x \to z = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_r]^{\mathrm{T}} \begin{bmatrix} K(x_1, x)) \\ \vdots \\ K(x_n, x)) \end{bmatrix} = \boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{K}(:, x) \tag{27}$$

式中,$\boldsymbol{K}(:, x) = [K(x_1, x), K(x_2, x), \cdots, K(x_n, x)]^{\mathrm{T}}$。

3　实验结果

这里,分别使用 UCI 数据集、文本数据集和图像数据集来验证文中所提出 SLDR 算法。为了评价新算法的性能,文中使用最近发表的 LDA-Km[21],SCREEN[1],SSDR[10] 和 LMDM[15] 与 SLDR 进行比较。由于 K 均值算法是一个简单有效的聚类算法,且在变换空间里,聚类性能充分反映维数约减算法的质量[1,7,15,18,21-22],因此,在投影空间里,文中使用 K 均值算法来聚类投影结果。进而,聚类结果作为最终衡量各算法性能的依据。

为了与几个流行算法作全面的对比,使用规范化精度度量来评价聚类结果[7,10,15,18]。给定聚类结果,精度度量表达如下

$$Accuracy = \sum_{i>j} \frac{\delta\{\delta\{c_i = c_j\} = \delta\{\hat{c}_i = \hat{c}_j\}\}}{0.5n(n-1)} \tag{28}$$

式中,$\delta\{\cdot\}$ 是指示函数,且 $\delta\{\mathrm{true}\} = 1, \delta\{\mathrm{false}\} =$

0。$\hat{c}_i$ 是数据 $\boldsymbol{x}_i$ 的聚类结果，$\boldsymbol{c}_i$ 是数据 $\boldsymbol{x}_i$ 的正确聚类标号。对任意一对样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，上述精度度量等价于由聚类算法得到的结果（$\hat{c}_i$，$\hat{c}_j$）与正确聚类标号（$\boldsymbol{c}_i$，$\boldsymbol{c}_j$）一致的概率。

在实验中，must-link 约束和 cannot-link 约束通过下面的方法得到。对于每个约束，随机地从输入样本集里选择一对样本，如果这对样本的类标号相同，则得到一个 must-link，否则得到一个 cannot-link。值得注意的是，这里使用的样本类标号是为了提供 must-link 约束、cannot-link 约束和比较各个算法性能的需要。在执行降维和聚类过程中，所有样本是没有类标号的。SLDR 算法在构造近邻图中，k 的取值为 5。为了公平比较，按照其他算法[1,10,15,21]的参数设置方法，聚类数取数据集的真正类数。同时，所有的算法运行在 MATLAB 环境里。5 个算法在每个数据集分别重复实验 40 次，取均值作为最终的聚类结果。

3.1 在 UCI 数据集上的实验

在本实验中，从 UCI 数据库中选择了 6 个数据集来比较 SLDR 算法与 LDA-Km，SCREEN，SSDR 和 LMDM 的性能，这些数据集分别是 Letter（a-d），Balance，Ionosphere，Segment，Vehicle 和 Soybean。为了研究不同成对约束的数量对各个算法性能的影响，在前 5 个数据集里，依次选择 40～400 个成对约束数，在 Soybean 数据集里，选择 10～100 个成对约束数。实验结果如图 1 所示，其中 n 是样本数，c 是聚类数，d 是原数据维数，r 是数据被降维后的数据维数。

从图 1 中发现，SLDR 算法在 Letter（a-d），Balance，Vehicle 和 Soybean 数据集上得到最好的聚类精度，在 Segment 和 Ionosphere 上的聚类精度也仅次于 SSDR 算法。这说明 SLDR 算法在减少数据维数时，能够找到较为满意的投影方向。

进一步，为了表明不同的数据维数对聚类精度的影响，我们在这 6 个数据集进行实验来比较 SLDR 算法与其他 4 个算法的性能，实验结果如图 2 所示（为了避免重复，我们只列出 5 个算法在 Letter（a-d），Soybean 和 Vehicle 上的实验结果，在其他数据集上，能得到与这 3 个数据集相似的观察）。

从图 2 中观察到，在数据维数发生变化时，SLDR 似乎总是能够得到最好的聚类结果，而且随着数据维数的变化，SLDR 的聚类精度并没有受到较大影响，这意味着新算法相当稳定。SSDR 与 SLDR 有相同的特性，但它的聚类精度要低于 SLDR。当数据维数发生变化时，SCREEN 和 LMDM 的聚类精度曲线有较大的波动，这意味着上述两个算法难以得到稳定的解。因此，联合图 1 和图 2，我们能得到初步的结论，SLDR 不仅得到较高的聚类精度，对数据维数的变化似乎也不敏感。

3.2 在文本数据集上的实验

在两个文本数据集 20Newsgroups 和 Reuters 上验证 5 个算法的性能。在 20Newsgroups 数据集里，选择 13 类 8 150 个样本，它们的维数是 2 500 维；在 Reuters 数据集中，选择 10 类 6 536 个样本，它们的维数是 1 500 维。实验环境和设置与前面的实验相同。

正如从图 3 中看到的那样，SLDR 得到最好的聚类精度。SSDR 尽管使用成对约束和无标号样本，但它的性能要差于 SLDR，主要原因是该算法忽略了数据的局部特性，这进一步证实了 Cai 和 Liu 等人的论证。SCREEN 和 LMDM 尽管比 LDA-Km 要好，但由于它们忽略了大量无标号样本对维数约减的贡献，因此，它们的精度要低于 SLDR 和 SSDR。在 5 个算法中，不论是在 UCI 数据集上，还是在文本数据集上，LDA-Km 的性能都最低，原因是该算法没有使用任何先验信息。这些结果告诉我们，在无监督算法中，一部分辅助信息的使用，确实能够提高其性能。

3.3 在人脸数据集上的实验

通过 3 个人脸数据集（ORL，PIE 和 YaleB）来研究 SLDR 算法的性能。ORL 数据集由不同光照、不同表情的 400 副人脸图像组成，其中每个人有 10 副图像。在本实验中，所取 ORL 中的人脸图像是大小 32×32 像素的 256 级灰度图像，如图 4 所示。并且选择 200 个成对约束，所有的样本都作为聚类对象。CMU PIE 数据集由 68 个人 41 368 副人脸图像组成。这些人脸图像包括不同光照、不同表情和不同姿态，如图 5 所示。同样，实验中所取的图像是大小 32×32 像素的 256 级灰度图像。选择 10 个人1 700 个人脸图像作为测试集，并选择 500 个成对约束。类似地，在 YaleB 中，选择 500 个成对约束，10 个人 3 850 个人脸图像作为测试集，如图 6 所示。为了公平比较各个算法，数据的维数被降到 $c-1$ 维（c 是聚类数），相应的实验结果如图 7 所示。

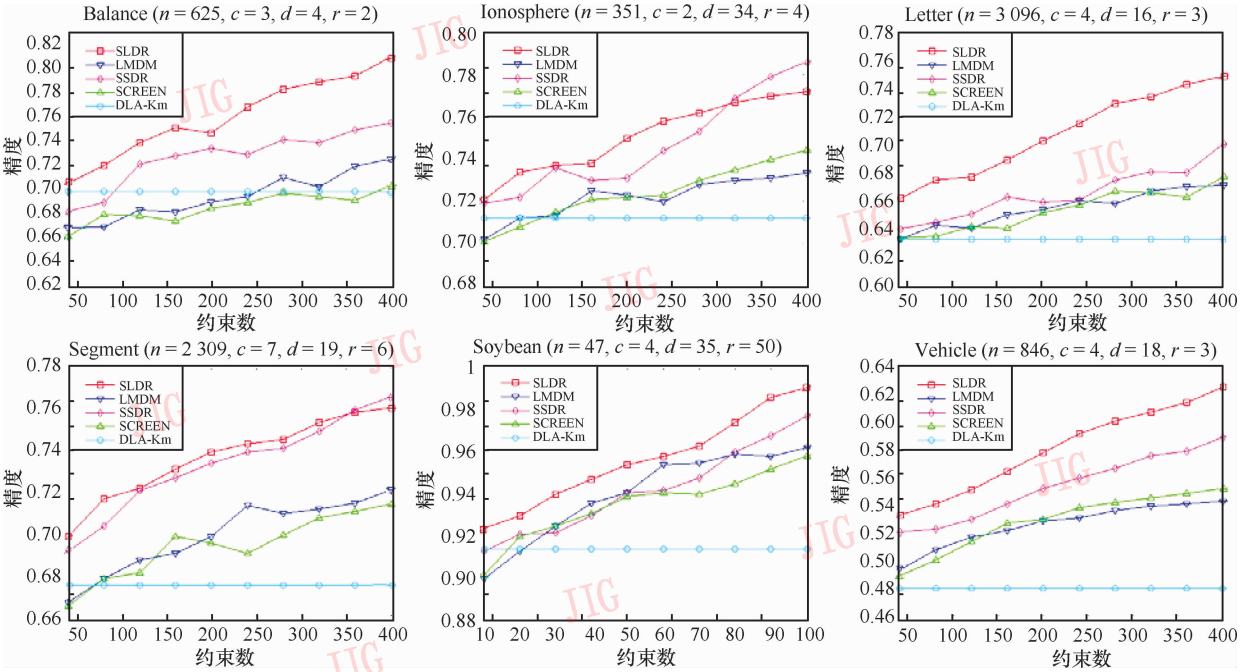

图 1　在 UCI 数据集上不同约束数的聚类精度

Fig. 1　Clustering accuracy on 6 UCI data sets with different number of constraints

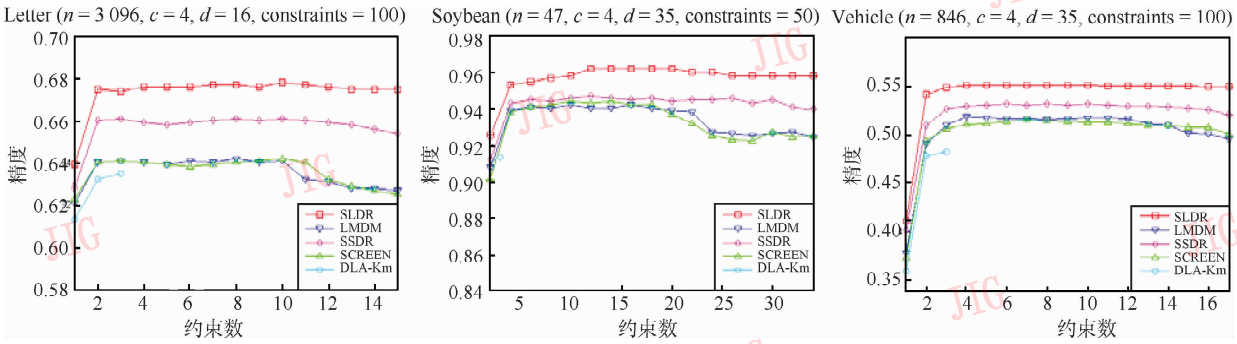

图 2　在 3 个 UCI 数据集上不同维数的聚类精度

Fig. 2　Clustering accuracy on 3 UCI data sets with different number of dimensions

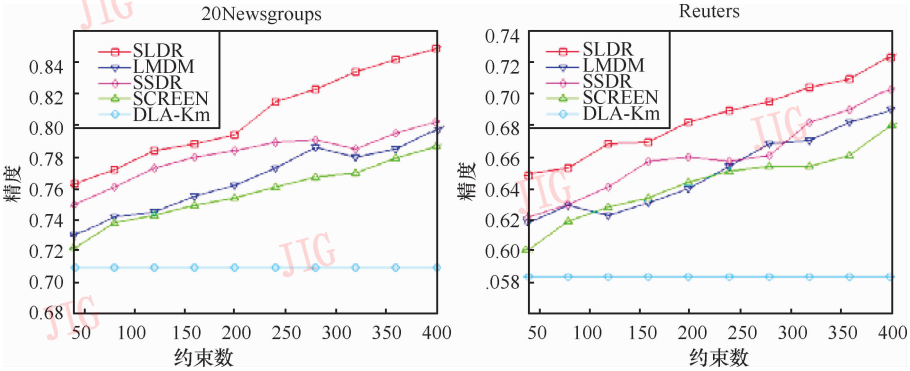

图 3　在两个文本数据集上不同约束数的聚类精度

Fig. 3　Clustering accuracy on 2 text data sets with different number of constraints

图 4 ORL 人脸数据集中的部分样本

Fig. 4 Sample face images from the ORL face database

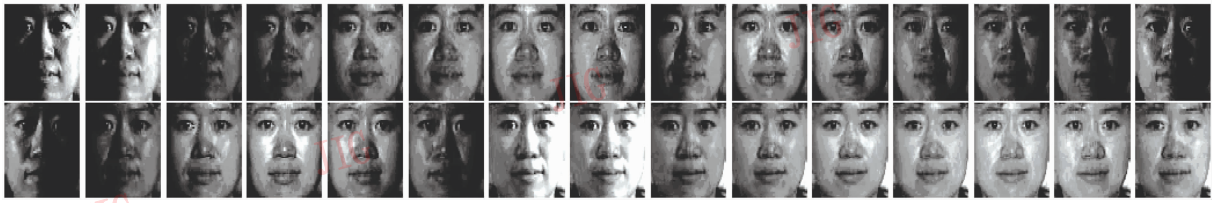

图 5 CMU PIE 人脸数据集中的部分样本

Fig. 5 Sample face images from the CMU PIE face database

图 6 YaleB 人脸数据集中的部分样本

Fig. 6 Sample face images from the Yale face B database

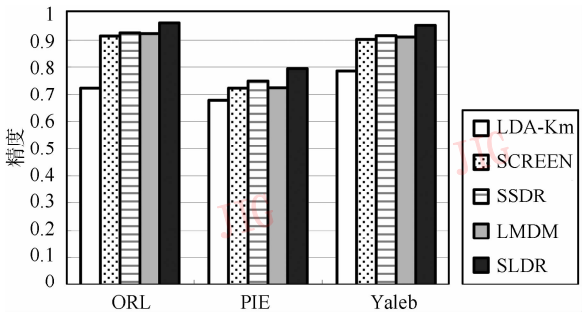

图 7 在 3 个人脸数据集上的聚类精度

Fig. 7 Clustering accuracy on 3 face data sets

3.4 讨 论

通过上面的 3 组实验,我们能得到如下几点观察。

1)SLDR 算法在大多数数据集上,包括 UCI 数据集、文本数据集和人脸数据集,无论是在少量约束数量还是大量约束数量条件下,都能够得到比 LDA-Km,SCREEN,SSDR 和 LMDM 更好的聚类精度。其中,在文本数据集和人脸数据集这些高维数据集上,SLDR 算法的优势更明显。因此,SLDR 算法是一个非常有效的半监督维数约减算法。

2)尽管在 Ionosphere 和 Segment 数据集上,SSDR 算法能得到最好的聚类精度,但在其他数据集上,该算法执行的效率都要比 SLDR 算法差,这是因为在先验信息不足的情况下对数据进行降维,数据的局部信息要比全局信息重要。另外,综合图 1—3,图 7 可以看到,SLDR 和 SSDR 比两个流行的半监督维数约减算法 SCREEN 和 LMDM 执行得好,原因是后两个算法仅仅使用成对约束信息,而忽略了大量无标号样本对维数约减的贡献。LDA-Km 算法是所有算法中执行的最差,这主要是该算法没有使用任何先验信息来引导维数约减过程。这也进一步验证了半监督算法要好于无监督算法。

3)随着成对约束数量的增加,SLDR 算法在大多数数据集上的聚类精度也逐渐增加,即新算法的曲线能保持上升趋势。而其他几个算法的曲线不能在数据上保持这种趋势。换句话说,随着成对约束数量的增加,它们的聚类精度反而有小幅下降。因此,SLDR 算法要比其他几个算法稳定。而且由于 SLDR 算法使用的是数据局部信息和先验信息来引导维数约减过程,以至于 SLDR 算法对数据中的噪声点并不敏感。

4　结　论

提出一种简单而有效的半监督局部维数约减算法(SLDR),该算法在使用大量无标号样本揭示数据几何结构的同时,使用专家提供的 must-link 和 cannot-link 成对约束来引导维数约减过程。另外,SLDR 算法较容易地推广到非线性空间,使之能够适应非线性数据的维数约减。在各种数据上的实验结果充分验证了所提出算法的有效性。

在下一步工作中,可研究一种方法(如主动学习的方法),找到最具有信息的成对约束,从而在成对约束不多的情况下,仍能使半监督维数约减方法得到令人满意的结果。

参考文献(References)

[1] Tang W, Hui X, Shi Z, et al. Enhancing semi-supervised clustering: a feature projection perspective[C]//Proceedings of the International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2007: 707-716.

[2] Saul L, Roweis S. Think globally. fit locally: Unsupervised learning of low dimensional manifolds[J]. Journal of Machine Learning Research, 2003, 4(1): 119-155.

[3] Tenenbaum J, de Silva V, Langford J. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[4] Jolliffe I. Principal Component Analysis[M]. New York: Springer, 2002.

[5] Bar-hillel A, Hertz T, Shental N, et al. Learning a mahalanobis metric from equivalence constraints[J]. Journal of Machine Learning Research, 2005, 6(5): 937-965.

[6] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]//Advances in Neural Information Processing Systems14. Cambridge, MA, USA: MIT Press, 2001: 585-591.

[7] Xing E, Ng A, Jordan M, et al. Distance metric learning, with application to clustering with side-information[C]//Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2003: 505-512.

[8] Bar-hillel A, Hertz T, Shental N, et al. Learning distance functions using equivalence relations[C]//Proceedings of the 20th International Conference on Machine Learning. Washington, DC, USA: 2003: 11-18.

[9] Cai D, He X, Han J. Semi-supervised discriminant analysis[C]//Proceedings of the 18th International Conference on Computer Vision. New York: ACM Press, 2007: 1-7.

[10] Zhang D, Zhou Z, Chen S. Semi-supervised dimensionality reduction[C]//Proceedings of SIAM Conference on Data Mining. New York: ACM Press, 2007: 629-634.

[11] Sugiyama M, Idé T, Nakajima S, et al. Semi-supervised local Fisher discriminant analysis for dimensionality reduction[C]//Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2008: 333-344.

[12] He X, Niyogi P. Locality preserving projections[C]//Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2003: 46-53.

[13] Song Y, Nie F, Zhang C, et al. A unified framework for semi-supervised dimensionality reduction[J]. Pattern Recognition, 2008, 41(9): 2789-2799.

[14] Sugiyama M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis[J]. Journal of Machine Learning Research, 2007, 8(5): 1027-1061.

[15] Xiang S, Nie F, Zhang C. Learning a Mahalanobis distance metric for data clustering and classification[J]. Pattern Recognition, 2008, 41(12): 3600-3612.

[16] Liu W, Tao D, Liu J. Transductive Component Analysis[C]//Proceedings of the International Conference on Data Mining. New York: ACM Press, 2008: 433-442.

[17] Cai D, He X, Zhou K, et al. Locality sensitive discriminant analysis[C]//Proceedings of International Joint Conference on Artificial Intelligence. Cambridge: MIT Press, 2007: 708-713.

[18] Yin Xuesong, Hu Enliang, Chen Songcan. Discriminative semi-

supervised clustering analysis with pairwise constraints[J]. Journal of Software, 2008, 19(11): 2791-2802. [尹学松, 胡恩良, 陈松灿. 基于成对约束的半监督判别分析[J]. 软件学报, 2008, 19(11): 2791-2802.]

[19] Zhao J, Lu K, He X. Locality sensitive semi-supervised feature selection[J]. Neurocomputing, 2008, 71(12): 1842-1849.

[20] Liu Y, Jin R, Jain A. BoostCluster: boosting clustering by pairwise constraints[C]//Proceedings of the International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 450-459.

[21] Ding C, Li T. Adaptive dimension reduction using discriminant analysis and K-means clustering[C]//Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2007: 521-528.

[22] Ye J, Zhao Z, Wu M. Discriminative K-Means for clustering[C]//Advances in Neural Information Processing Systems 20. Cambridge, MA, USA: MIT Press, 2007: 1649-1656.

[23] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]//Proceedings of the 18th International Conference on Machine Learning. New York: ACM Press, 2001: 577-584.

[24] Xu Sen, Lu Zhimao, Gu Guochang. Two spectral algorithms for ensembling document clusters[J]. Acta Automatica Singica, 2009, 35(7): 997-1002. [徐森, 卢志茂, 顾国昌. 解决文本聚类集成问题的两个谱算法[J]. 自动化学报, 2009, 35(7): 997-1002.]