软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(1):28−43 [doi: 10.3724/SP.J.1001.2011.03928]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

半监督降维方法的实验比较*

陈诗国, 张道强+

(南京航空航天大学 计算机科学与工程系,江苏 南京 210016)

Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods

CHEN Shi-Guo, ZHANG Dao-Qiang+

(Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

+ Corresponding author: E-mail: dqzhang@nuaa.edu.cn

Chen SG, Zhang DQ. Experimental comparisons of semi-supervised dimensional reduction methods. *Journal of Software*, 2011,22(1):28−43. http://www.jos.org.cn/1000-9825/3928.htm

Abstract: Semi-Supervised learning is one of the hottest research topics in the technological community, which has been developed from the original semi-supervised classification and semi-supervised clustering to the semi-supervised regression and semi-supervised dimensionality reduction, etc. At present, there have been several excellent surveys on semi-supervised classification: Semi-Supervised clustering and semi-supervised regression, e.g. Zhu's semi-supervised learning literature survey. Dimensionality reduction is one of the key issues in machine learning, pattern recognition, and other related fields. Recently, a lot of research has been done to integrate the idea of semi-supervised learning into dimensionality reduction, i.e. semi-supervised dimensionality reduction. In this paper, the current semi-supervised dimensionality reduction methods are reviewed, and their performances are evaluated through extensive experiments on a large number of benchmark datasets, from which some empirical insights can be obtained.

Key words: semi-supervised dimensionality reduction; dimensionality reduction; semi-supervised learning; class label; pairwise constraint

摘 要: 半监督学习是近年来机器学习领域中的研究热点之一,已从最初的半监督分类和半监督聚类拓展到半监督回归和半监督降维等领域.目前,有关半监督分类、聚类和回归等方面的工作已经有了很好的综述,如 Zhu 的半监督学习文献综述.降维一直是机器学习和模式识别等相关领域的重要研究课题,近年来出现了很多将半监督思想用于降维,即半监督降维方面的工作.有鉴于此,试图对目前已有的一些半监督降维方法进行综述,然后在大量的标准数据集上对这些方法的性能进行实验比较,并据此得出了一些经验性的启示.

关键词: 半监督降维;降维;半监督学习;类别标号;成对约束

中图法分类号: TP181 **文献标识码**: A

在很多机器学习和模式识别的实际应用中,人们经常会遇到高维数据,如人脸图像、基因表达数据、文本

* 基金项目: 国家自然科学基金(60875030); 模式识别国家重点实验室开放课题(20090044)

收稿时间: 2009-12-18; 定稿时间: 2010-07-28

CNKI 网络优先出版: 2010-11-05 11:50, http://www.cnki.net/kcms/detail/11-2560.TP.20101105.1150.001.html

© 中国科学院软件研究所 http://www.jos.org.cn

数据等.直接对这些高维数据进行处理是非常费时且费力的,而且由于高维数据空间的特点,容易出现所谓的“维数灾难”问题[1].降维(dimensionality reduction)是根据某一准则,将高维数据变换到有意义的低维表示[2].因此,降维能够在某种意义上克服维数灾难.根据是否使用了类别标号,传统的降维方法可以分为两类:无监督降维,如主成分分析(PCA)[3];有监督降维,如线性判别分析(LDA),或称为 Fisher 判别分析(FDA)[4].

在很多实际任务中,无标号的数据往往很容易获取,而有标号的数据则很难获取.为了获得更好的学习精度同时又要充分地利用现有的数据,出现了一种新的学习形式,即半监督学习(semi-supervised learning).相比较传统的学习方法,半监督学习可以同时利用无标号数据和有标号数据,只需较少的人工参与就能获得更精确的学习精度,因此,无论在理论上,还是在实践中,都受到越来越多的关注.目前,半监督学习已从最初的半监督分类和半监督聚类拓展到半监督回归和半监督降维等领域.有关半监督分类、聚类和回归等方面的工作已经有了很好的综述,如 Zhu 的半监督学习文献综述[5].然而据我们所知,目前国内外尚未专门针对半监督降维方面的综述工作.因此,本文试图对目前已有的一些半监督降维方法进行综述,然后在大量的标准数据集上比较这些方法的性能,并根据实验得出一些经验性的启示.

本文第 1 节介绍降维的概念,并对已有的降维方法进行分类.第 2 节详细介绍当今较为流行的几种典型的半监督降维方法,本文将这些方法大致分为 3 类:基于类别标号的方法、基于成对约束的方法和基于其他监督信息的方法.第 3 节分别在 UCI 标准数据集、半监督学习数据集[6]和标准人脸数据上比较第 2 节中介绍的半监督降维方法,并作一些经验性的讨论.最后一节对本文的工作进行总结并指出下一步研究的方向.

1 降 维

给定一批观察样本,记 $X\in R^{D\times N}$,包含 N 个样本,每个样本有 D 个特征.降维的目标是:根据某个准则,找到数据的低维表示 $Z=\{z_i\}\in R^d(d<D)$,同时保持数据的“内在信息(intrinsic information)”[3].当降维方法为线性时,降维的过程就转变为学习一个投影矩阵 $W=\{w_i\}_{i=1}^d\in R^{D\times d}$,使得

$$Z=W^TX \tag{1}$$

其中,T 表示矩阵的转置操作.当降维方法为非线性时,不需要学习这样一个投影矩阵 W,而直接从原始数据中学习得到低维的数据表示 Z.

图 1 对当今流行的一些降维方法进行了分类.首先,根据是否使用数据中的监督信息,将所有方法分成监督的(supervised)、半监督的(semi-supervised)和无监督的(unsupervised)降维这样 3 类.其中,根据监督信息的不同,半监督降维又可分为基于类别标号(class label)的、基于成对约束(pairwise constraints)的和基于其他监督信息的 3 类方法.然后再根据算法模型的不同,又可以将所有的降维方法分成线性(linear)降维和非线性(nonlinear)降维.最后列出了一些有代表性的降维方法以及该方法所出自的文献.下面对图中出现的每一种降维方法进行简要的说明.

线性判别分析(LDA)也叫 Fisher 判别分析(FDA)[4],是当今最流行的监督降维方法之一.其主要思想是,寻找一个投影矩阵,使得降维之后同类数据之间尽量紧凑,而不同类别数据之间尽量分离.Baudat 和 Anouar 使用核技巧,把 LDA 扩展到非线性形式,即广义判别分析(GDA)[7].间隔 Fisher 判别分析(MFA)[8]和局部 Fisher 判别分析(LFDA)[9]是传统 FDA 的两个扩展版本.与上面的方法不同,判别成分分析(DCA)[10]是利用成对约束进行度量学习(成对约束的定义将在第 2.2 节中给出介绍).其主要思想与 LDA 类似:寻找一个投影矩阵,使得降维之后正约束数据之间尽量紧凑,而负约束数据之间尽量分离.KDCA 是它的核化版本[10].

上述降维方法都是有监督的,需要知道数据的某种监督信息,如类别标号或者成对约束等.无监督降维则不需要这些监督信息,它直接利用无标号的数据,在降维过程中保持数据的某种结构信息.主成分分析(PCA)[3]是一种典型的无监督降维方法,其目的是寻找在最小平方意义下最能够代表原始数据的投影[1].与 PCA 不同,多维尺度分析(MDS)[11]使得变换后的低维数据点之间的欧氏距离与原数据点之间的欧氏距离尽量保持一致.非负矩阵分解(NMF)[12]则基于这样的假设:数据矩阵可以分解为两个非负矩阵的乘积——基矩阵和系数矩阵.核 PCA(KPCA)[13]是传统 PCA 方法的核化版本.KPCA 中的核函数需要人为地指定,而最大方差展开(MVU)[14]则通

© 中国科学院软件研究所 http://www.jos.org.cn

过对数据的学习直接得到核矩阵.除了上面提到的一些非监督降维方法以外,流形学习(manifold learning)是最近发展起来的一种新的降维方法,它假设数据采样于高维空间中的一个潜在流形上,通过寻找这样一个潜在的流形很自然地找到高维数据的低维表示.ISOMAP[15]、局部线性嵌入(LLE)[16]、拉普拉斯特征映射(LE)[17]和局部保持投影(LPP)[18]是流形学习的代表性方法.

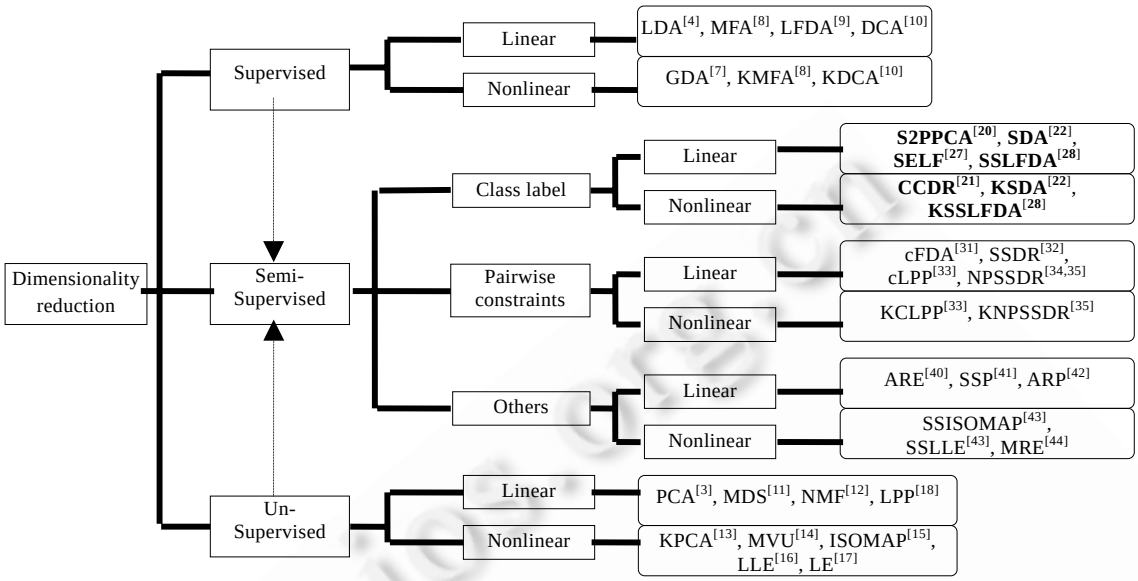


Fig.1 Taxonomy of dimensionality reduction methods

图 1 降维方法的分类图

上面分别介绍了监督降维和无监督降维的一些具有代表性的方法,下一节重点介绍半监督降维以及一些典型的半监督降维方法.

2 半监督降维

把半监督学习思想用于降维,就形成了半监督学习的一个新的分支,即半监督降维.半监督降维是传统降维方法的有效综合,它既可以像监督降维方法那样利用数据标号,又可以像无监督降维方法那样保持数据的某种结构信息,如数据的全局方差、局部结构等等.因此,半监督降维能够克服传统降维方法的缺点,有重要的研究价值和广阔的应用前景.根据使用监督信息的不同,半监督降维方法可以大致分成 3 类:(1) 基于类别标号的方法;(2) 基于成对约束的方法;(3) 基于其他监督信息的方法.

2.1 基于类别标号的半监督降维

首先,我们给出基于类别标号半监督降维方法的数学描述.假设有 N 个数据 $X=\{x\}_{i=1}^{N}$,每个数据的维数为 D,即 $x_i\in R^D(i=1,2,\ldots,N)$.在所有数据中,$L$ 个数据已经知道类别标号,记为 $X_1=\{(x_i,y_i)\}_{i=1}^{L}$,其中,$x_i$ 表示第 i 个数据,y_i 是 x_i 的类别标号,总共有 C 个类;剩下的数据没有类别标号,记为 $X_2=\{x_j\}_{j=L+1}^{N}$.半监督降维的目的是,利用有类别标号的数据和无类别标号的数据 $X=\{X_1,X_2\}$,寻找数据的低维表示 $Z=\{z_i\}\in R^d(d<D)$.

近年来,研究者们已经提出了多种基于类别标号的半监督降维方法.Yu 等人在概率 PCA 模型[19]的基础上加入了类别标号信息,提出了监督和半监督形式的概率 PCA 模型[20].Costa 和 Hero 在构造拉普拉斯图时引入了类别标号信息,得到了拉普拉斯特征映射算法的一种监督和半监督版本[21].Cai 等人在传统的 LDA 方法中加入流形正则化项,提出了一种半监督的判别分析方法 SDA[22].Song 等人提出了一个半监督降维方法框架[23],SDA 可以看成是该框架下的一个例子.Zhang 等人在文献[24]中提出的半监督降维方法与 SDA 方法相似,也是使用

© 中国科学院软件研究所 http://www.jos.org.cn

正则化项来保持数据的流形结构.所不同的是,它使用了一种基于路径的鲁棒的相似性来构造邻接图.文献[25]在最大化 LDA 准则中加入没有类别标号的数据,使用约束凹凸过程解决最终的优化问题.Chen 等人把 LDA 重写成最小平方的形式,通过加入拉普拉斯正则化项,该模型可以转化为一个正则化的最小平方问题[26].最近,Sugiyama 把局部 Fisher 判别分析(LFDA)和 PCA 结合起来,提出了一种半监督局部降维 Fisher 判别分析 SELF[27].Chatpatanasiri 等人从流形学习的角度提出了一个半监督降维框架.在该框架下,可以很容易地把传统的 Fisher 判别分析扩展到半监督的形式[28].

下面我们简要介绍 5 种半监督降维方法:半监督概率 PCA(S2PPCA)[19]、分类约束降维(CCDR)[21]、半监督判别分析(SDA)[23]和两个半监督 Fisher 判别分析 SELF[27],SSLFDA[28].最后,对这几种方法的属性以及它们之间的关系进行分析.

2.1.1 半监督概率 PCA(S2PPCA)

S2PPCA 是概率 PCA 模型[20]的半监督版本.首先,仅考虑有标号的样本 X_1.假设样本(x,y)由下列隐变量模型生成:

$$x=W_x z+\mu_x+\varepsilon_x,\ y=f(z,\Theta)+\varepsilon_y \tag{2}$$

这里,$f(z,\Theta)=[f_1(z,\theta_1),\ldots,f_c(z,\theta_C)]^T$ 是关于类别标号的函数,其中 $\Theta=\{\theta_1,\ldots,\theta_C\}$表示 C 个确定性函数 $f_1,\ldots,f_C$ 的参数(假设每个函数 $f_c,c=1,\ldots,C$,都是关于隐变量 z 的线性函数 $f_c(z,\theta_c)=w_y^{cT}z+\mu_y^c$,则 $f(z,\Theta)=W_x z+\mu_y$).$z\sim N(0,I)$. $z\sim N(0,I)$是输入 x 与输出 y 所共享的隐变量.两个相互独立的噪声模型被定义成各向同性的高斯函数,即 $\varepsilon_x\sim N(0,\delta_x^2 I),\varepsilon_x\sim N(0,\delta_x^2 I)$.因此,对隐变量 z 求积分,得到样本(x,y)的似然函数:

$$P(x,y)=\int P(x,y\mid z)P(z)\mathrm{d}z=\int P(x\mid z)P(y\mid z)P(z)\mathrm{d}z \tag{3}$$

这里,$x\mid z\sim N(W_x z+\mu_x,\theta_x^2 I),y\mid z\sim N(W_y z+\mu_y,\theta_y^2 I)$.如果样本之间相互独立,则有 $P(X_1)=\prod_{i=1}^{L}P(x_i,y_i)$.最后,所有需要估计的参数向量表示为 $\Omega=\{W_x,W_y,\mu_x,\mu_y,\delta_x^2,\delta_y^2\}$.然后,考虑所有样本 $X=\{X_1,X_2\}$的情况.因为样本之间假设是相互独立的,那么关于有所有样本的似然函数为

$$P(X)=P(X_1)P(X_2)=\prod_{i=1}^{L}P(x_i,y_i)\prod_{j=L+1}^{N}P(x_j) \tag{4}$$

这里,$P(x_i,y_i)$可以由公式(3)计算得到,而 $P(x_j)=\int P(x_j\mid z_j)P(z_j)\mathrm{d}z_j$ 可以由概率 PCA 模型计算得到.

2.1.2 分类约束降维(CCDR)

CCDR 的主要思想如下:将每个类所有样本的中心点作为新的数据节点加入到邻接图中,然后同类样本点与它们的中心点之间加入一条权重为 1 的边.这样,CCDR 可以形式化地写成最小化下面的目标函数:

$$E(Z_n)=\sum_{ki}a_{ki}\|z_k-y_i\|^2+\beta\sum_{ij}w_{ij}\|y_i-y_j\|^2 \tag{5}$$

这里,z_k表示嵌入到低维空间后第 k 个类的中心;$A=\{a_{ki}\}$表示类别关系矩阵(如果数据 x_i 属于第 k 类,则 $a_{ik}=1$;否则,$a_{ik}=0$);$W=\{w_{ij}\}$表示数据的邻接图,它的构造方法有很多种,可以参考文献[29];y_i 是指嵌入到低维空间后的数据向量,其中,$Z_n=[z_1,\ldots,z_C,y_1,\ldots,y_N]$.

2.1.3 半监督判别分析(SDA)

SDA 是基于线性判别分析的一个半监督降维版本,它通过在 LDA 的目标函数中添加正则化项,使得 SDA 在最大化类间离散度的同时可以保持数据的局部结构信息.SDA 能够优化下面的目标函数:

$$\arg\max_{w}\frac{w^T S_b w}{w^T S_t w+\alpha J(w)}=\arg\max_{a}\frac{w^T S_b w}{w^T(S_t+\alpha XLX^T+\beta I)w} \tag{6}$$

式中,S_b 表示带标号数据的类间离散度矩阵,S_t 表示总体离散度,$J(W)$是正则化项(通过构造 k-近邻图保持数据的流形),L 表示拉普拉斯矩阵.像 LDA 一样,SDA 的目标函数也可以转化为一个广义特征分解问题.

2.1.4 半监督局部 Fisher 判别分析(SELF)

SELF 是由文献[27]提出来的另一种半监督降维方法,它是局部 Fisher 判别分析[9]的半监督版本.SELF 通过

© 中国科学院软件研究所 http://www.jos.org.cn

把 PCA 和 LFDA 综合起来,可以保持无类别标号数据的全局结构,同时保留 LFDA 方法的优点(比如类内的数据为多模态分布、LDA 的维数限制等).SELF 可以表示为求解下面的优化问题:

$$Wopt = \arg\max_{W}[tr(W^T S^{(rlb)} W(W^T S^{(rlw)} W)^{-1})] \tag{7}$$

上式中,W 是映射矩阵.$S^{(rlb)}$是正则化局部类间离散矩阵,$S^{(rlw)}$是正则化局部类内离散矩阵,定义如下:

$$S^{(rlb)}=(1-\beta)S^{(lb)}+\beta S^{(t)},\ S^{(rlw)}=(1-\beta)S^{(lw)}+\beta I_d \tag{8}$$

其中,$S^{(lb)}$和 $S^{(lw)}$分别是 LFDA 算法中的局部类间离散矩阵和局部类内离散矩阵,$S^{(t)}$是离散度矩阵(数据方差矩阵).$\beta\in[0,1]$是模型的调节参数,当β=1 时,SELF 就退化为 PCA;而当β=0 时,SELF 就退化为 LFDA.

2.1.5 基于流形学习的半监督局部 Fisher 判别分析(SSLFDA)

与第 2.1.4 节中的 SELF 算法类似,SSLFDA 也是 LFDA 算法的一个半监督版本,它是根据文献[28]中提出的半监督降维框架直接推导得到的.它与 SELF 在利用无类别标号数据方面有所不同:SSLFDA 算法保持数据的流形结构,而 SELF 保持数据的全局结构.在这个框架里面,半监督降维方法可以简单地表示为求解下面的优化问题:

$$W^* = \arg\min_{W} f^l(W^T X) + \gamma f^u(W^T X) \tag{9}$$

其中,$f^l(\cdot)$和 $f^u(\cdot)$分别表示关于有类别数据和无类别标号数据的函数,γ是调节因子.通常,f 可以写成成对数据加权距离的函数,最终,问题(9)转化为矩阵的特征分解问题.通过定义不同的权值,该框架可以导出不同的半监督降维方法.

2.1.6 方法总结

表 1 中列出了第 2.1.1 节~第 2.1.5 节中 5 个半监督降维方法的一些属性.k 表示数据邻接图近邻样本点的数目,t 是高斯核函数(Gaussian-kernel)中的带宽,i 表示迭代的次数,p 表示稀疏矩阵中的非零元素的个数.因为在 S2PPCA 方法中提供了两个互为对偶的 EM 算法,所以表中列出了它的两个计算复杂度和存储复杂度.

S2PPCA 是概率模型,算法的性能一方面依赖于模型假设,即数据分布是高斯的或混合高斯的;另一方面还依赖于样本数据的个数.如果样本数目太少,参数的估计就不可信,从而导致算法的性能下降.CCDR 是一种非线性方法,与 LE 方法一样,邻接图构造的好坏会直接影响算法的性能.SDA,SELF 和 SSLFDA 是 3 种线性降维方法.SDA 通过增加正则化项,使得在降维的过程中能够保持数据的局部结构;SELF 需要最大化数据的协方差(PCA 准则),因此它在降维的过程中利用无标号的数据保持数据的全局结构;SSLFDA 利用无标号的数据保持数据的流形结构(LPP 准则),使数据的局部结构得到保持.

Table 1 Properties of semi-supervised dimensionality reduction methods based on class label

表 1 基于类别标号的半监督降维方法的属性

Methods	Basic idea	Parameters	Computational	Memory
S2PPCA	Label+PPCA	None	$O(i(D+L)Nd)$ or $O(N^2(id+D))$	$O((D+L)N)$ or $O(N^2)$
CCDR	Label+LE	k, t, β	$O(p(N+C)^2)$	$O(p(N+C)^2)$
SDA	LDA+Adjacency graph	k, α, β	$O(D^3)$	$O(N^2+D^2)$
SELF	LFDA+PCA	k, β	$O(D^3)$	$O(D^2)$
SSLFDA	LFDA+LPP	k, β	$O(D^3)$	$O(D^2)$

2.2 基于成对约束的半监督降维

在半监督学习中,除了类别标号信息,还可以利用其他形式的先验信息,比如成对约束.在很多情况下,人们往往不知道样本的具体类别标号,却知道两个样本属于同一个类别,或者不属于同一个类别,我们称这样的监督信息为成对约束.成对约束往往分为两种:正约束(must-link)和负约束(cannot-link).正约束指的是两个样本属于同一个类别;相反地,负约束指的是两个样本属于不同的类别.本文中,把所有正约束的集合记为 ML,所有负约束的集合记为 CL.

我们首先回顾一些基于成对约束的半监督降维方法.Tang 等人提出用约束指导降维过程[30],他们的方法仅仅用到约束而忽略了无标号的数据.Bar-Hillel 等人提出一种约束 FDA(cFDA)[31]方法对数据进行预处理,它是

© 中国科学院软件研究所 http://www.jos.org.cn

作为相关成分分析算法(RCA)的一个中间步骤.Zhang 等人从一个更为直观的角度同时利用正约束和负约束指导降维过程,提出了一个半监督降维框架(SSDR)[32].Cevikalp 等人在局部保持投影方法(LPP)中引入约束信息,提出了约束局部保持投影算法(cLPP)[33].Wei 等人提出了一种邻居保持降维(NPSSDR)方法[34],在利用约束指导降维的同时,保持数据的局部结构信息.Baghshah 等人将 NPSSDR 方法用于度量学习[35],他们使用了一种二分搜索方法来优化求解过程.Chen 等人在文献[36]中提出了一个基于约束信息的半监督非负矩阵分解(NMF)框架.彭岩等人在传统的典型相关分析算法中加入成对约束,提出了一种半监督典型相关分析算法[37].最近,Davidson 提出了一个基于图的降维框架[38],在该框架中,首先构造一个约束图,然后根据构建出来的图来指导降维.

下面我们简要介绍 4 种基于成对约束的半监督降维方法:约束 Fisher 判别分析(cFDA)[31]、基于约束的半监督降维框架(SSDR)[32]、约束的局部保持投影(cLPP)[33]和邻域保持半监督降维(NPSSDR)[34,35].最后,对这几种方法的属性以及它们之间的关系进行分析.

2.2.1 约束 Fisher 判别分析(cFDA)

cFDA 是度量学习算法相关成分分析(RCA)[31]的一个中间步骤.cFDA 的具体做法是:首先,使用正约束(文献[31]中称为 equivalence constraints)把数据聚成若干个簇(cluster);然后,类似于 LDA 构建簇内散布矩阵 S_w 和总体散布矩阵 S_t;最后,最大化下面的比率:

$$\max_{W}\frac{W^T S_t W}{W^T S_w W} \tag{10}$$

其中,W 是映射矩阵,T 是矩阵转置符号.优化目标 W 可以简单地由矩阵 $S_w^{-1}S_t$ 的前 d 个特征向量组成.

2.2.2 基于约束的半监督降维框架(SSDR)

不同于 cFDA 利用约束信息来构造散布矩阵,SSDR[32]直接使用约束来指导降维.SSDR 在降维的过程中保持数据之间的约束关系,同时像 PCA 一样保持数据内部结构信息.SSDR 最大化下面的目标方程:

$$J(w)=\frac{1}{2n^2}\sum_{i,j}(w^T x_i-w^T x_j)^2+\frac{\alpha}{2n_{CL}}\sum_{(x_i,x_j)\in CL}(w^T x_i-w^T x_j)^2-\frac{\beta}{2n_{ML}}\sum_{(x_i,x_j)\in ML}(w^T x_i-w^T x_j)^2 \tag{11}$$

上式中,第 1 项可以使得降维后两两数据之间的距离保持最大,它等价于 PCA 准则,即最大化数据的方差;n_{CL} 和 n_{ML} 分别表示负约束和正约束的个数.

2.2.3 约束的局部保持投影(cLPP)

与 SSDR 不同,cLPP 在降维过程中保持数据的局部结构信息.cLPP 的具体步骤如下:首先,构造数据的邻接矩阵;然后,利用约束信息修改邻接矩阵中的权值,使得正约束数据之间的权值增大,负约束数据之间的权值变小,同时修改与有约束的数据点直接相连点的相关权值,对约束信息进行传播;最后,cLPP 的目标函数可以显示地写成下面的形式:

$$J(w)=\frac{1}{2}\left(\sum_{i,j}(z_i-z_j)^2\tilde{A}_{ij}+\sum_{i,j\in ML}(z_i-z_j)^2-\sum_{i,j\in CL}(z_i-z_j)^2\right) \tag{12}$$

这里,$\tilde{A}_{ij}$ 代表修改后的数据邻接矩阵,z_i 是原始数据 x_i 映射到低维空间后所对应的点.

2.2.4 邻域保持半监督降维(NPSSDR)**

NPSSDR 利用正约束和负约束降维,同时保持数据的局部结构信息.与 cLPP 方法不同,NPSSDR 不需要构造数据的邻接矩阵,而是通过添加正则化项的方法来实现.这里,我们给出了文献[35]中的目标函数:

$$W^*=\arg\max_{W^TW=I}\frac{\sum_{(z_i,z_j)\in CL}(z_i-z_j)^2}{\sum_{(z_i,z_j)\in ML}(z_i-z_j)^2+\alpha J(W)} \tag{13}$$

** Wei 等人在文献[34]中提出 NPSSDR 方法.Baghshah 等人在文献[35]中提出的算法思想与文献[34]相似,我们统一称他们的方法为 NPSSDR.实验中比较的 NPSSDR 方法是按照文献[35]中的求解方法实现的.

© 中国科学院软件研究所 http://www.jos.org.cn

式中,$J(W)$是正则化项.若使用局部线性嵌入(LLE)的思想构造正则化项,则有 $J(W)=tr(W^TXMX^TW)$,其中,M 是数据重构矩阵.于是,目标函数(13)最终转化为求解一个迹比(trace radio)问题,该问题可以使用一种二分搜索算法来求解[39].

2.2.5 方法总结

表 2 中显示了基于成对约束的半监督降维方法的一些属性,k 表示数据邻接图近邻样本点的数目,t 是高斯核函数中的带宽,i 表示迭代的次数,p 表示稀疏矩阵中的非零元素的个数.

cFDA 仅仅利用正约束而没有用负约束,它对约束的选取有很强的依赖性;并且,当 S_W 奇异的时候,cFDA 的求解过程还需要作特殊的处理.SSDR 既可以利用正约束,也可以利用负约束,但是它在降维过程中只保持数据的全局结构,并且约束的信息也没有进行传播.cLPP 与 SSDR 相比,保持了数据的局部结构,并且把约束信息传播到邻近的数据点.但是,cLPP 与 LPP 方法一样,降维性能的好坏需要依赖于数据邻接图的构建.NPSSDR 使用一种二分搜索算法近似求解最终的问题,这种搜索算法需要反复求解特征值问题,计算复杂度较大,并且算法有可能会遇到不收敛的问题.NPSSDR 使用 LLE 策略保持无标号数据的局部结构,它会遇到 LLE 同样的问题,比如局部结构塌陷问题等等[2].

Table 2 Properties of semi-supervised dimensionality reduction methods based on pairwise constraints

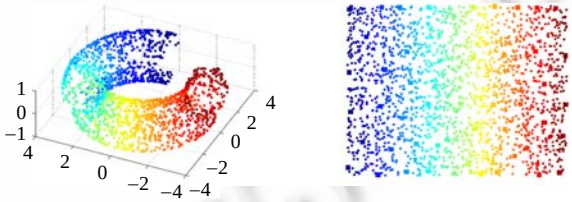
表 2 基于成对约束的半监督降维方法的属性

Methods	Basic idea	Structure	Parameters	Computational	Memory
cFDA	must-link+LDA	Global	None	$O(D^3)$	$O(D^2)$
SSDR	cannot-link+must-link+PCA	Global	α, β	$O(D^3)$	$O(N^2+D^2)$
cLPP	cannot-link+must-link+LPP	Local	k, t	$O(D^3)$	$O(N^2+D^2)$
NPSSDR	cannot-link+must-link+LLE	Local	k, α	$O(iD^3+pN^2)$	$O(D^2+pN^2)$

2.3 基于其他监督信息的方法

半监督降维方法除了可以利用类别标号和成对约束作为监督信息以外,还有很多其他形式的监督信息,我们把它们统一分在了第 3 类.扩充关系嵌入(ARE)[40]、语义子空间映射(SSP)[41]和相关集成映射(RAP)[42]是利用图像检索中的检索与被检索图像间的相关关系作为监督信息指导特征抽取的过程.Yang 等人使用流形上的嵌入关系,把一些无监督的流形方法扩展到半监督的形式[43],例如半监督的 ISOMAP(SS-ISOMAP)和半监督的局部线性嵌入(SS-LLE).Memisevic 等人提出了一种半监督降维框架多关系嵌入(MRE)[44],可以综合利用多种相似性关系.

图 2 展示了一个使用流形上的嵌入关系的例子[43].图中相对较大的实心样本点表示已知嵌入关系的样本,即在嵌入过程中,这些样本在低维空间中的位置是已知的,而其他样本的位置是未知的.文献[43]就是利用这种已知的嵌入关系作为先验知识,把几种经典的流形学习方法扩展到半监督的形式.



(a) Original samples (b) Low dimensional embedded samples[43]

(a) 原始数据 (b) 低维嵌入[43]

Fig.2 Prior information in the form of on-manifold coordinates

图 2 已知低维嵌入关系作为先验知识

3 实 验

在这一节中,我们在大量的标准数据集上比较了第 2 节中介绍的几种半监督降维方法.这些标准数据集包

© 中国科学院软件研究所 http://www.jos.org.cn

括 UCI 标准数据集***、半监督学习数据集[5]和人脸标准数据集.实验中,对于基于类别标号的降维方法,我们按照不同的比例将数据集随机划分为两个部分:一部分(占数据总数的 10%,30%或 50%)作为有标号的数据,另外一部分作为无标号的数据.对于基于成对约束的降维方法,随机选取一定数目(占数据总数的 10%,30%或 50%)的成对约束,而把整个数据集当作无标号的数据.对于 UCI 标准数据集和半监督数据集,选用最大似然估计方法[45]计算数据集的内在维数,文献[2]中的实验中也采用了相同的处理方法.实验采用简单的最近邻分类器(nearest neighbor,简称 NN)的分类精度作为降维方法的评价指标,使用留一交叉验证法估计最终的实验结果.各种降维方法的参数设置见表 3,除了部分参数经验设定以外,其他参数都采用网格式搜索的方法确定最优设置.

Table 3 Parameter settings for experiments

表 3 实验中的参数设置

Methods	Parameter settings
PCA	None
LDA	None
S2PPCA	None
CCDR	$1\leqslant k\leqslant 15$, $t=1$, $0<\beta\leqslant 10$
SDA	$1\leqslant k\leqslant 15$, $\beta=0$, $0<\alpha\leqslant 10$
SELF	$1\leqslant k\leqslant 15$, $0<\beta<1$
SSLFDA	$1\leqslant k\leqslant 15$, $0<\gamma\leqslant 10$, $\alpha=8$
cFDA	None
DCA	None
SSDR	$\alpha=1$, $1\leqslant\beta\leqslant 30$
cLPP	$1\leqslant k\leqslant 15$, $t=5$
NPSSDR	$1\leqslant k\leqslant 15$, $\alpha=0.2$ or 0.02

3.1 UCI标准数据集上的实验比较

本实验中使用了 8 个 UCI 标准数据集用于测试降维算法的性能.表 4 中列出了这些数据集的属性(C 表示类别数目,D 表示维数,N 表示样本数目).这些数据集中既包括简单的数据集,例如 iris 和 soybean_small,也包括一些复杂的数据集,例如手写数据集 digits0.05 和 letter0.05.

Table 4 Properties of 8 UCI datasets

表 4 8 个 UCI 数据集的属性

	iris	digits0.05	letter0.05	protein	soybean_small	letter_0.1_IJL	ionosphere	zoo
C	3	10	26	6	4	3	2	7
D	4	16	16	20	35	16	34	16
N	150	550	1 000	116	47	227	351	101

表 5 和表 6 分别显示了基于类别标号和成对约束的半监督降维方法在 UCI 标准数据集上的实验结果.表中列出来的结果是经过降维后的数据再用最近邻分类器得到的分类精度,每个实验结果都是算法各自运行 20 遍后的平均值.PCA,LDA 和只利用约束的 DCA 作为比较的基准方法也被列了出来.表中的最左列表示的是数据集的名称、最大似然估计得到的维数值 D 和类别数 C.表 5 中的 nL 表示的是有标号的数据在整个数据集中所占的比例(10%,30%或 50%),表 6 中的 nC 表示的是约束的数目占数据总数的比例(10%,30%或 50%),其中,正约束和负约束的数目相同.黑色粗体数值表示的是该数据集上最好的实验结果.

*** 实验中使用的 UCI 标准数据集可以从网站 http://archive.ics.uci.edu/ml/上得到,其中,digits0.05,letter0.05 和 letter_0.1_IJL 是从原始数据集 digits,letter 中进行抽样得到的.具体的抽样方法可参照工具包 Weka(http://www.cs.waikato.ac.nz/ml/weka/)中的数据集.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 5 Results of different methods based on class label on UCI datasets

表 5 基于类别标签半监督降维方法在 UCI 标准数据集上的实验结果

Dataset	*nL*	PCA	LDA	S2PPCA	CCDR	SDA	SELF	SSFLDA
iris (2*D*, 3*C*)	10%	0.960 0	0.955 7	0.908 3	0.896 3	**0.969 3**	0.951 0	0.967 0
	30%	0.960 0	0.966 7	0.943 7	0.904 3	0.967 7	0.954 7	**0.970 0**
	50%	0.960 0	0.964 3	0.964 7	0.889 3	**0.966 7**	0.955 0	0.957
digits0.05 (4*D*, 10*C*)	10%	0.869 1	0.812 8	**0.891 5**	0.296 8	0.885 5	0.878 2	0.881 8
	30%	0.869 1	0.851 3	0.898 2	0.461 2	0.893	**0.899 1**	0.894 9
	50%	0.869 1	0.860 7	**0.923 4**	0.597 1	0.893 6	0.898 2	0.895 0
letter0.05 (5*D*, 26*C*)	10%	0.527 0	0.471 4	0.594 1	0.231 9	0.646 3	0.637 7	**0.657 1**
	30%	0.527 0	0.626 6	0.668 4	0.399 2	0.672 6	0.674 1	**0.675 9**
	50%	0.527 0	0.639 7	**0.739 8**	0.567 5	0.668 6	0.685 2	0.684 0
protein (8*D*, 6*C*)	10%	0.732 8	0.604 3	0.687 9	0.502 2	0.681 9	**0.748 3**	0.695 7
	30%	0.732 8	0.728 4	0.753 4	0.548 3	0.749 1	0.771 1	**0.783 6**
	50%	0.732 8	0.792 7	0.805 6	0.643 1	0.772 4	**0.813 4**	0.807 3
soybean_small (2*D*, 4*C*)	10%	0.957 4	0.661 7	0.935 1	0.672 3	0.784 0	0.941 5	**0.959 6**
	30%	0.957 4	0.787 2	0.936 2	0.812 8	0.713 8	0.966 0	**0.967 0**
	50%	0.957 4	0.851 1	0.945 7	0.811 7	0.652 1	0.983 0	**0.986 2**
letter_0.1_ijl (4*D*, 3*C*)	10%	0.859 0	0.738 8	0.890 3	0.695 2	0.909 5	0.898 7	**0.927 5**
	30%	0.859 0	0.895 6	0.912 1	0.672 5	0.921 8	0.915 9	**0.932 6**
	50%	0.859 0	0.909 9	0.934 6	0.729 3	0.925 3	0.918 9	**0.938 6**
ionosphere (11*D*, 2*C*)	10%	0.903 1	0.712 4	0.844 3	**0.953 7**	0.814 8	0.873 5	0.903 4
	30%	0.903 1	0.737 2	0.868 2	**0.959 3**	0.835 0	0.887 3	0.908 6
	50%	0.903 1	0.752 4	0.881 5	**0.968 4**	0.835 9	0.887 0	0.909 0
zoo (3*D*, 7*C*)	10%	0.960 4	0.707 9	0.941 1	0.799 5	**0.964 4**	0.962 4	0.948 5
	30%	0.960 4	0.868 8	0.921 3	0.914 9	**0.969 8**	0.963 9	0.960 9
	50%	0.960 4	0.883 7	0.924 8	0.934 2	**0.975 7**	0.964 4	0.963 4

Table 6 Results of different methods based on pairwise constraints on UCI datasets

表 6 基于成对约束半监督降维方法在 UCI 标准数据集上的实验结果

Dataset	*nC*	PCA	DCA	cFLD	SSDR	cLPP	LPSSDR
iris (2*D*, 3*C*)	10%	**0.960 0**	0.941 3	0.948 3	0.953 7	0.959 7	0.938 3
	30%	0.960 0	0.961 0	0.952 0	0.962 3	**0.967 7**	0.932 3
	50%	0.960 0	0.967 7	0.959 0	0.960 0	**0.973 7**	0.930 3
digits0.05 (4*D*, 10*C*)	10%	0.869 1	0.484 7	0.694 6	**0.888 4**	0.868 1	0.861 0
	30%	0.869 1	0.675 0	0.668 7	**0.878 1**	0.863 6	0.864 3
	50%	0.869 1	0.627 3	0.603 1	**0.878 1**	0.850 6	0.857 6
letter0.05 (5*D*, 26*C*)	10%	0.527 0	0.319 6	0.496 2	0.548 4	**0.628 7**	0.595 6
	30%	0.527 0	0.451 0	0.470 6	0.559 4	**0.662 2**	0.637 0
	50%	0.527 0	0.433 6	0.424 7	0.556 0	**0.669 4**	0.649 0
protein (8*D*, 6*C*)	10%	**0.732 8**	0.594 8	0.664 7	0.725 0	0.719 8	0.699 1
	30%	0.732 8	0.678 9	0.723 7	0.757 3	0.746 6	**0.759 9**
	50%	0.732 8	0.750 0	0.735 3	0.771 1	0.754 7	**0.775 9**
soybean_small (2*D*, 4*C*)	10%	**0.957 4**	0.833 0	0.839 4	0.893 6	0.624 5	0.884 0
	30%	0.957 4	0.864 9	0.893 6	0.966 0	0.586 2	**0.991 5**
	50%	0.957 4	0.984 0	0.709 6	**0.995 7**	0.626 6	0.992 6
letter_0.1_IJL (4*D*, 3*C*)	10%	0.859 0	0.774 0	0.789 0	0.917 0	**0.927 1**	0.891 9
	30%	0.859 0	0.837 4	0.758 2	0.910 4	**0.934 4**	0.897 8
	50%	0.859 0	0.720 7	0.684 6	0.906 0	**0.933 7**	0.907 3
ionosphere (11*D*, 2*C*)	10%	0.903 1	0.716 4	0.827 9	**0.909 5**	0.891 0	0.891 0
	30%	0.903 1	0.866 7	0.814 7	**0.904 3**	0.901 7	0.901 7
	50%	0.903 1	0.845 9	0.809 1	**0.915 5**	0.863 5	0.912 4
zoo (3*D*, 7*C*)	10%	0.960 4	0.865 8	0.884 7	**0.961 4**	0.945 5	0.943 6
	30%	0.960 4	0.904 0	0.867 8	**0.971 8**	0.963 4	0.951 0
	50%	0.960 4	0.868 8	0.833 2	**0.973 8**	0.962 9	0.966 3

下面是表 5 和表 6 中实验结果的总结和简要分析:第一,一些半监督降维方法在 UCI 数据集上的性能确实

© 中国科学院软件研究所 http://www.jos.org.cn

比 PCA 要好.例如,基于类别标号的方法 SSLFDA 和基于成对约束的方法 SSDR.LDA 只能利用有类别标号的数据,当训练数据较少时会存在过拟合的危险,所以降维性能甚至还不如 PCA.而半监督降维方法既可以使用类别标号信息,又可以使用无标号的数据,可以充分利用数据中的信息,所以降维性能常常好于传统的方法;第二,观察每种方法的实验结果,随着类别标号或者约束数目的增加,半监督降维方法的性能通常都有一定程度的提高.这个现象证实了我们的直觉,如果可以提供足够多的先验信息,或者最具有代表性的先验信息,就可以在一定程度上提升降维性能;第三,有些半监督降维方法(例如 CCDR)在某些数据集上的性能并不理想.我们分析出现这种情况的原因可能是:(1) CCDR 的求解过程与 LE 是一样的,因此它可能会遇到与 LE 方法相同的问题,比如特征分解时的数值问题[2];(2) CCDR 目标函数比较简单,只有在有标号数据足够多[21]或者数据类别数目很少的时候(实验中数据集 ionosphere),它才能获得良好的性能.总之,一些半监督降维方法在 UCI 数据集上表现出较传统方法更好的性能,但是我们并不能从实验结果中明显地看出哪一种方法比其他方法更好,这些半监督方法性能的优劣在很大程度上取决于具体的数据集.因此,在设计新的半监督降维方法时,需要根据具体的应用环境选择合适的算法模型,而寻找一种在所有数据下都能获得最优结果的方法是不现实的.

3.2 半监督学习数据集上的实验比较

半监督学习数据集是文献[6]中提供的一组标准数据集(共有 9 组,分别是 Set1,Set2,…,Set9),用于比较半监督学习算法的性能.每个数据集都提供了两组数据的划分(10 和 100 个数据有类别标号)用于产生有标号的数据和无标号的数据,每组数据各有 12 种具体的划分结果.其中,Set8 样本数目太多,而 Set3 和 Set6 采用 PCA 降维后得到分类精度已经很高,因此,本文只选用剩下的 6 个数据集.包含 3 个人工数据集和 3 个真实数据集.表 7 列出了这 6 个数据集的具体属性(C 表示类别数目,D 表示维数,N 表示样本数目).

Table 7 Properties of semi-supervised learning dataset

表 7 半监督学习数据集的属性

	Set1 (Digit1)	Set5 (g241c)	Set7 (g241n)	Set2 (USPS)	Set4 (BCI)	Set9 (Text)
C	2	2	2	2	2	2
D	241	241	241	241	117	11 960
N	1 500	1 500	1 500	1 500	400	1 500
Comment	Artificial	Artificial	Artificial	Handwritten digits	Brain computer interface	Text

表 8 和表 9 分别显示了基于类别标号和成对约束的半监督降维方法在半监督数据集上的实验结果.表中列出来的是经过降维后的数据再用最近邻分类器得到的分类精度,每个实验结果都是算法各自运行 12 遍后的平均值(根据数据集的划分).实验中,成对约束是由有标号的数据部分随机产生的,对每个数据集分别产生 12 组成对约束,每一组都含有相同数目的正约束和负约束(100 或 10).PCA,LDA 和 DCA 作为实验比较的基准方法.表中的最左列表示的是数据集的名称、最大似然估计得到的维数值 D 和数据集的类别数 C.表 8 中的 nL 表示的是有标号的数据的个数,表 9 中的 nC 表示的是正约束和负约束的个数.黑体数值表示的是该数据集上最好的实验结果.

Table 8 Results of different methods based on class label on semi-supervised learning dataset

表 8 基于类别标号半监督降维方法在半监督数据集上的实验结果

Dataset	nL	PCA	LDA	S2PPCA	CCDR	SDA	SELF	SSFLDA
Set1 (Digit1)	10	0.979 3	0.530 2	0.965 1	**0.985 9**	0.742 3	0.978 8	0.923 1
(41D, 2C)	100	0.979 3	0.527	0.885 3	**0.984 7**	0.905 6	0.981 2	0.979 4
Set2 (USPS)	10	**0.970 7**	0.694 1	0.939 3	0.911 2	0.762	0.967 4	0.868 8
(22D, 2C)	100	0.970 7	0.712 2	0.938 9	0.901 6	0.839 3	**0.974 3**	0.958 6
Set4 (BCI)	10	0.512 5	0.503 3	0.539 6	0.557 7	0.528 5	0.555 6	**0.562 1**
(9D, 2C)	100	0.512 5	0.611 5	0.589 2	0.646 0	0.657 3	**0.737 7**	0.702 1
Set5 (g241c)	10	0.642 0	0.500 3	0.612 8	0.514 9	0.521 6	**0.708 7**	0.501 9
(70D, 2C)	100	0.642 0	0.539 7	0.656 5	0.534 4	0.621 1	**0.755 9**	0.507 2
Set7 (g241n)	10	0.721 3	0.503 1	0.669	0.514 3	0.523 6	**0.770 4**	0.512 2
(68D, 2C)	100	0.721 3	0.540 0	0.676 4	0.530 7	0.610 1	**0.772 8**	0.505 6
Set9 (Text)	10	0.657 3	0.512 4	0.513 8	**0.804 3**	0.524 8	0.559 2	0.615 2
(40D, 2C)	100	0.657 3	0.536 7	0.553 4	**0.802 3**	0.562 7	0.567 9	0.621 9

© 中国科学院软件研究所 http://www.jos.org.cn

Table 9 Results of different methods based on pairwise constraints on semi-supervised learning dataset

表 9 基于成对约束的半监督降维方法在半监督数据集上的实验结果

Dataset	*nC*	PCA	DCA	cFLD	SSDR	cLPP	LPSSDR
Set1 (Digit1)	10	0.979 3	0.832 2	0.978 8	0.980 0	**0.982 7**	0.974 5
(41*D*, 2*C*)	100	0.979 3	0.908 5	0.972 1	0.979 1	**0.981 3**	0.976 1
Set2 (USPS)	10	0.970 7	0.963 0	0.963 0	0.966 2	**0.975**	0.924
(22*D*, 2*C*)	100	0.970 7	0.862 3	0.870 3	**0.970 5**	0.965 6	0.952 2
Set4 (BCI)	10	0.512 5	0.535	**0.566 3**	0.565 2	0.492 5	0.539 2
(9*D*, 2*C*)	100	0.512 5	0.517 3	0.513 3	0.536 9	**0.679 2**	0.555 4
Set5 (g241c)	10	0.642 0	0.545 2	**0.684 4**	0.656 0	0.593 5	0.577 5
(70*D*, 2*C*)	100	0.642 0	0.548 8	0.551 6	**0.661 1**	0.578 4	0.629 7
Set7 (g241n)	10	0.721 3	0.606 3	**0.798 4**	0.715 3	0.641 8	0.598 3
(68*D*, 2*C*)	100	**0.721 3**	0.543 6	0.620 8	0.701 8	0.588 7	0.634 4
Set9 (Text)	10	0.657 3	0.559 1	0.646 4	0.584 4	**0.677 2**	0.622 8
(40*D*, 2*C*)	100	0.657 3	0.540 0	0.652 1	0.564 5	**0.685 3**	0.642 2

下面分别对两组实验结果进行一些经验性的总结和分析.

(1) 一些在 UCI 标准数据集上表现很好的方法在半监督学习数据集上的性能并不是最好的.例如,基于类别标号的方法 SSLFDA 和基于成对约束的方法 SSDR.这是因为,与 UCI 标准数据集相比,半监督学习数据集的数据分布更加复杂,而同时提供的先验信息(类别标号或成对约束)却更少;

(2) 表 8 中的方法 SELF 在大多数数据集上的结果都很好,而在 Set9 上的结果较差,而表 9 中的方法 cLPP 在 Set1,Set2 和 Set9 上的结果很好,而在 Set5 和 Set7 上的结果不如其他一些方法.这个现象说明,方法 SELF 和方法 cLPP 的性能在很大程度上依赖于具有特定性质的数据集.

总之,实验结果进一步说明,没有最好的降维方法,只有适合于具体数据的降维方法.比较第 3.1 节和第 3.2 节中的实验,对于数据分布更加复杂的半监督数据集,半监督降维方法对数据的依赖性表现得更为突出,即在有的数据集上性能较好而在另一些数据集上性能却变差.同时,半监督方法往往需要调节更多的参数,这极大地增加了半监督降维方法应用于大规模数据集的难度.

3.3 在人脸标准数据集的实验比较

实验在 3 个人脸标准数据集上验证了半监督降维方法的性能.这 3 个人脸标准数据集分别是 AR 数据集、PIE 数据集和 YALE 数据集[****].其中:AR 数据集包含 100 个人的总共 1 400 幅脸部正面图片,每张图片的大小重新调整为 33×24;PIE 数据集包含 68 个人的脸部正面图片部分,每张图片的大小重新调整为 64×64;YALE 数据集包含 15 个人的总共 165 幅图片,每张图片的大小重新调整为 64×64.图 3~图 5 给出了半监督降维方法在这 3 个数据集上的实验结果.图中横坐标表示的是降维后的数据的维数,纵坐标表示的是最近邻分类器在降维后数据上的分类精度,每个实验结果都是运行 10 次后的平均值.“30%(或 50%) labels”表示有类别标号数据的数目占总数的 30%(或 50%).类似地,“30%(或 50%) constraints”表示成对约束的数目是数据总数的 30%(或 50%).我们没有比较 S2PPCA 方法,因为在这些人脸数据集上,S2PPCA 方法的收敛速度很慢.实验中,我们对数据集作了预处理:首先使用 PCA 把数据降到较低的维数(PCA 率为 0.98),然后再使用监督或半监督降维方法作进一步的处理.下面,我们归纳了图 3~图 5 的一些现象,并对其进行简要的分析.

(1) 在人脸图像上,大多数半监督降维方法的性能都好于 PCA;基于类别标号的半监督降维方法的性能好于 LDA(例如 SDA),或者与 LDA 差不多(例如 SSLFDA 和 SELF);大多数基于成对约束的半监督降维方法的性能都好于监督方法 DCA;

(2) 半监督降维方法 CCDR 的性能不如 PCA(数据集 AR),或者只是在维数较少的时候才好于 PCA(数据集 PIE 和 YALE).我们认为其原因在于:一方面,CCDR 需要计算最小特征值,在数值计算中会遇到琐碎解问题[2];另

**** AR 数据集的下载地址是 http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html.PIE 数据集的下载地址:http://www.ri.cmu.edu/research project_detail.html?project_id=418&menu_id=261.YALE 数据集的下载地址是 http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

© 中国科学院软件研究所 http://www.jos.org.cn

一方面,表示流形结果所需要的样本个数往往会随着数据的内在维数呈指数倍增长,即流形嵌入时发生“维数灾难”[2];

(3) 在数据集 PIE 和 YALE 上,DCA 的实验曲线在上升了一点之后往往不会再发生变化,这可能是因为:DCA 在求解目标方程的过程中舍弃了零特征值所对应的特征向量,使得目标维数的个数受到了限制;

(4) 比较图中的几种半监督降维方法我们可以看出,SDA 和 SSDR 比同类(基于类别标号或者基于成对约束)的方法性能更优.而更为复杂的方法,例如 SSLFDA 和 cLPP 反而没有这两种简单方法的效果好,这似乎在某种程度上说明了简单模型的重要性;

(5) 在 PIE 和 YALE 数据集上,基于约束的半监督降维方法的实验结果一般都呈现出随着维数的增加先上升然后再下降的趋势.这个现象的可能原因是:随着维数的增加,数据结构将变得更加复杂,更为简单的成对约束则不能够提供足够的判别信息来指导降维.

综合以上结果,与第 3.1 节和第 3.2 节的实验中找不出最好的降维方法的现象不同,在人脸数据集上,SDA 和 SSDR 比其他半监督方法表现出更好的性能.我们认为,出现这种差别的原因在于:实验中所用的 UCI 标准数据集和半监督数据集都是由不同性质的数据集组成的,这些数据集的分布特征可能是多种多样的,而第 3.3 节中所用的 3 个人脸标准数据集都采样于同类数据源(人脸),这些数据在分布上可能具有某种相似的性质,而这两种方法正好适用于具有这种性质的数据,因此在实验中能够获得较好的性能.

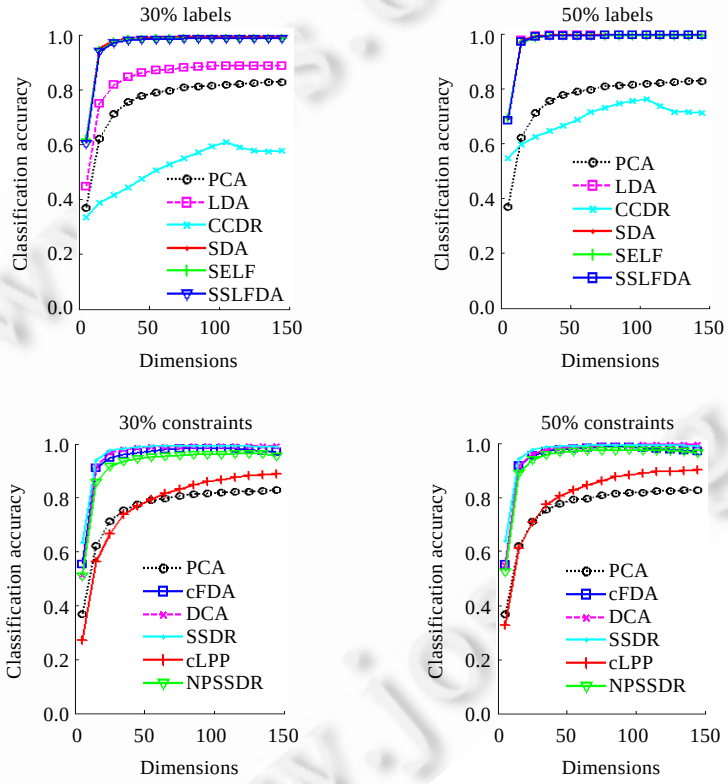


Fig.3 Results on AR

图 3 AR 数据集上的实验结果

© 中国科学院软件研究所 http://www.jos.org.cn

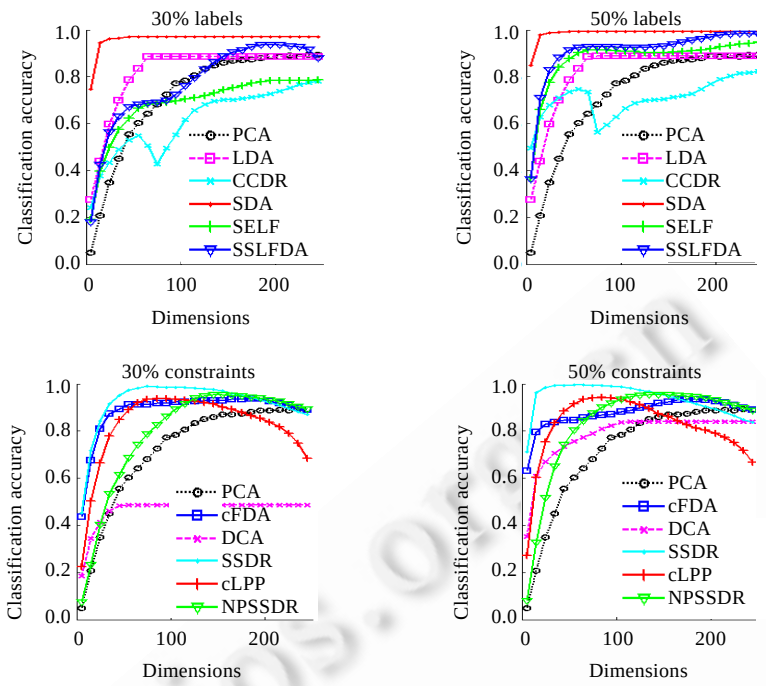


Fig.4 Results on PIE

图 4 PIE 数据集上的实验结果

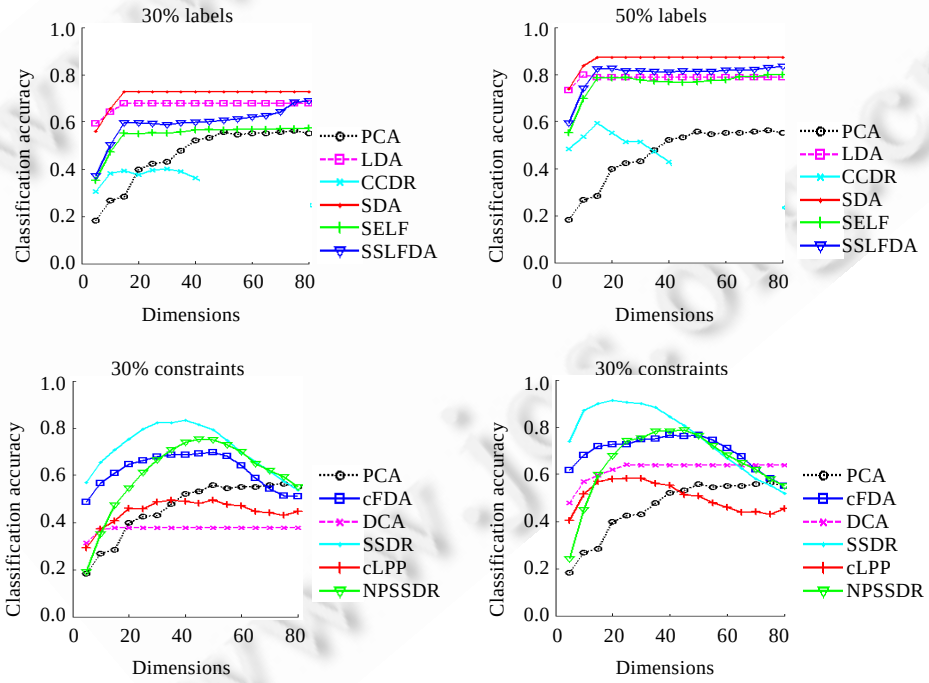


Fig.5 Results on YALE

图 5 YALE 数据集上的实验结果

© 中国科学院软件研究所 http://www.jos.org.cn

4 总 结

本文首先回顾了近些年来提出的一些半监督降维方法,然后分 3 组实验比较了这些方法和传统的监督、无监督降维方法在大量数据集上的性能,并对实验结果进行了一些经验性的探讨和分析.限于篇幅,本文的实验只比较了一些线性半监督降维方法(CCDR 除外)在分类中的性能.未来的工作包括:(1) 比较这些半监督降维方法在聚类和数据可视化方面的性能;(2) 比较非线性半监督降维方法在更加复杂的数据集(如图像、文本)上的性能.另外,本文中的半监督降维方法在利用无标号数据时,要么保持数据的局部结构,要么保持数据的全局结构(如最大数据方差),而最近的研究[46,47]表明,保持数据的稀疏性结构可以有效提升分类或降维的性能.因此,我们下一步的工作还包括把稀疏性思想引入到半监督降维中.

另一方面,虽然半监督降维方法在大多数数据集上都表现出良好的性能,但是还有一些重要的问题没有解决,归纳如下:

(1) 数据内在维数的估计.高维数据(比如人脸图像)的内在维数往往并不是很高的,若能准确估计得到数据的内在维数,许多半监督降维方法的目标维数就不再需要人为地指定.然而,当数据十分复杂或者非线性时,估计数据的内在维数是一个非常困难的问题;

(2) 监督信息的选择.在实际的学习任务中,获取监督信息通常需要付出很大的代价.因此,如何选择最具信息量的监督信息就非常重要.以分类为例,直观上,位于类与类之间的数据对判别来说更有信息含量.若能有效选择这一部分的数据并对其进行标号,将会在很大程度上提升半监督降维算法的性能;

(3) 正约束和负约束的选择.对于基于成对约束的半监督降维方法,选择不同的正约束和负约束的比例,同样会影响降维的性能.目前已有研究者指出,正约束可能比负约束含有更多的信息[31,32],但是尚缺乏足够令人信服的理由;

(4) 评价指标.本文中,我们选择分类精度作为降维方法的评价指标.但是,降维的目的不仅仅局限于分类任务,它还有其他很多形式的应用,比如聚类和数据可视化.选择一个合适的评价指标,是正确评价算法性能的关键性因素之一.

References:

[1] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed., New York: John Wiley & Sons, 2001. 170.

[2] van der Maaten L, Postma E, van den Herik J. Dimension reduction: A comparative review. Technical Report, TiCC-TR 2009-005, Tilburg University, 2009.

[3] Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 1933,24(6):417−441. [doi: 10.1037/h0071325]

[4] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936,7(2):179−188.

[5] Zhu X. Semi-Supervised learning literature survey. Technical Report, Computer Sciences TR 1530, Department of Computer Sciences, University of Wisconsin at Madison, 2008.

[6] Chapelle O, Scholkopf B, Zien A, eds. Semi-Supervised Learning. Cambridge: MIT Press, 2006.

[7] Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. Neural Computation, 2000,12(10):2385−2404. [doi: 10.1162/089976600300014980]

[8] Yan S, Xu D, Zhang B, Zhang HJ. Graph embedding and extensions a general framework for dimensionality reduction. In: Proc. of the CVPR 2005. San Diego, 2005. [doi: 10.1109/CVPR.2005.170]

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

[10] Hoi S, Liu W, Lyu MR. Learning distance metrics with contextual constraints for image retrieval. In: Proc. of the CVPR 2006. New York: IEEE Computer Society, 2006. 2076−2078. [doi:10.1109/CVPR.2006.167]

[11] Torgerson WS. Multidimensional scaling: Theory and method. Psychometrika, 1952,17(4):401−419. [doi: 10.1007/BF02288916]

[12] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001. http://books.nips.cc/papers/files/nips13/LeeSeung.pdf

© 中国科学院软件研究所 http://www.jos.org.cn

[13] Scholkopf B, Smola AJ, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5):1299−1319. [doi: 10.1162/089976698300017467]

[14] Weinberger KQ, Sha F, Saul LK. Learning a kernel matrix for nonlinear dimensionality reduction. In: Proc. of the ICML 2006. Pittsburgh, 2006. [doi: 10.1145/1015330.1015345]

[15] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500):2319−2323. [doi: 10.1126/science.290.5500.2319]

[16] Roweis ST, Saul LK. Nonlnear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326.

[17] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, Vol.17. Cambridge: MIT Press, 2003. http://books.nips.cc/papers/ files/ nips14/AA42.pdf

[18] He X, Niyogi P. Locality preserving projections. In: Thrun S, Saul L, Schölkopf B, eds. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2003. http://books.nips.cc/papers/files/nips16/NIPS2003_AA20.pdf

[19] Tipping E, Bishop CM. Probabilistic principal component analysis. Journal of the Royal Statistical Society, 1999,B(61):611−622.

[20] Yu S, Yu K, Tresp V, Kriegel HP, Wu M. Supervised probabilistic principal component analysis. In: Proc. of the KDD 2006. New York: ACM, 2006. 464−473. [doi:10.1145/1150402.1150454]

[21] Costa JA, Hero AO. Classification constrained dimensionality reduction. In: Proc. of the ICASSP 2005, Vol.5. Philadelphia, 2005. 1077−1080. [doi: 10.1109/ICASSP.2005.1416494]

[22] Cai D, He X, Han J. Semi-Supervised discriminant analysis. In: Proc. of the ICCV 2007. Rio de Janeiro, 2007. 1−7. [doi: 10.1109/ICCV.2007.4408856]

[23] Song Y, Nie F, Zhang C, Xiang S. A unified framework for semi-supervised dimensionality reduction. Pattern Recognition, 2008, 41(9):2789−2799. [doi: 10.1016/j.patcog.2008.01.001]

[24] Zhang Y, Yeung D. Simi-Supervised discriminant analysis using robust path-based similarity. In: Proc. of the CVPR 2008. Anchorage, 2008. 1−8. [doi: 10.1109/CVPR.2008.4587357]

[25] Zhang Y, Yeung D. Semi-Supervised discriminant analysis via CCCP. In: Proc. of the ECML PKDD 2008. Berlin, Heidelberg: Springer-Verlag, 2008. 644−659. [doi: 10.1007/978-3-540-87481-2_42]

[26] Chen J, Ye J, Li Q. Integrating global and local structures: A least squares framework for dimensionality reduction. In: Proc. of the CVPR 2007. Minneapolis, 2007. 1−8. [doi: 10.1109/CVPR.2007.383040]

[27] Sugiyama M, Ide T, Nakajima S, Sese J. Semi-Supervised local fisher discriminant analysis for dimensionality reduction. Machine Learning, 2008,78(1-2):35−61. [doi: 10.1007/s10994-009-5125-7]

[28] Chatpatanasiri R, Kijsirikul B. A unified semi-supervised dimensionality reduction framework for manifold learning. Neurocomputing, 2010,73(10-12):1631−1640. [doi: 10.1016/j.neucom.2009.10.024]

[29] Cai D, He X, Han J. Document clustering using locality preserving indexing. IEEE Trans. on Knowledge and Data Engineering, 2005,17(12):1624−1637.

[30] Tang W, Zhong S. Pairwise constraints-guided dimensionality reduction. In: Proc. of the Workshop on Feature Selection for Data Mining (SDM 2006). Bethesda, 2006. http://www.siam.org/meetings/sdm06/workproceed/FSDM/FSDM-proceedings2006.pdf#page=64

[31] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2006,6(6):937−965.

[32] Zhang D, Zhou Z, Chen S. Semi-Supervised dimensionality reduction. In: Proc. of the SDM 2007. Minneapolis, 2007. 629−634. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.92.8730

[33] Cevikalp H, Verbeek J, Jurie F, Klaser A. Semi-Supervised dimensionality reduction using pairwise equivalence constraints. In: Proc. of the VISAPP 2008. Funchal, 2008. 489−496. http://eprints.pascal-network.org/archive/00003909/01/Constrained_Clustering.pdf.

[34] Wei J, Peng H. Neighbourhood preserving based semi-supervised dimensionality reduction. Electronics Letters, 2008,44(20): 1190−1191.

[35] Baghshah MS, Shouraki SB. Semi-Supervised metric learning using pairwise constraints. In: Proc. of the IJCAI 2009. San Francisco: Morgan Kaufmann Publishers, 2009. 1217−1222.

© 中国科学院软件研究所 http://www.jos.org.cn

[36] Chen Y, Rege M, Dong M, Hua J. Incorporating user provided constraints into document clustering. In: Proc. of the ICDM 2007. Washington: IEEE Computer Society, 2007. 103−112. [doi: 10.1109/ICDM.2007.67]

[37] Peng Y,Zhang DQ. Semi-Supervised canonical correlation analysis algorithm. Journal of Software, 2008,19(11):2822−2832 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2822.htm [doi: 10.3724/SP.J.1001.2008.02822]

[38] Davidson I. Knowledge driven dimension reduction for clustering. In: Proc. of the IJCAI 2009. San Francisco: Morgan Kaufmann Publishers, 2009. 1034−1039.

[39] Xiang S, Nie F, Zhang C. Learning a mahalanobis distance metric for data clustering and classification. Pattern Recognition, 2008,41(12):3600−3612. [doi: 10.1016/j.patcog.2008.05.018]

[40] Lin Y, Liu T, Chen H. Semantic manifold learning for image retrieval. In: Proc. of the MM 2005. New York: ACM, 2005. 249−258. [doi: 10.1145/1101149.1101193]

[41] Yu J, Tian Q. Learning image manifolds by semantic subspace projection. In: Proc. of the MM 2006. New York: ACM, 2006. 297−306. [doi: 10.1145/1180639.1180710]

[42] Liu W, Jiang W, Chang SF. Relevance aggregation projections for image retrieval. In: Proc. of the CIVR 2008. New York: ACM, 2008. 119−126. [doi: 10.1145/1386352.1386372]

[43] Yang X, Fu H, Zha H, Barlow J. Semi-Supervised nonlinear dimensionality reduction. In: Proc. of the ICML 2006. New York: ACM, 2006. 1065−1072. [doi: 10.1145/1143844.1143978]

[44] Memisevic R, Hinton G. Multiple relational embedding. In: Saul L, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems, Vol.17. Cambridge: MIT Press, 2004. http://books.nips.cc/papers/files/nips17/NIPS2004_0835.pdf

[45] Levina E, Bickel PJ. Maximum likelihood estimation of intrinsic dimension. In: Saul L, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems, Vol.17. Cambridge: MIT Press, 2004. http://books.nips.cc/papers/files/ nips17/ NIPS2004_0094.pdf

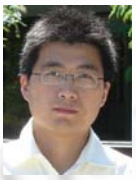
[46] Wright J, Yang A, Sastry S, Ma Y. Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(2):210−227. [doi: 10.1109/TPAMI.2008.79]

[47] Qiao L, Chen S, Tan X. Sparsity preserving projections with applications to face recognition. Pattern Recognition, 2010,43(1): 331−341. [doi: 10.1016/j.patcog.2009.05.005]

附中文参考文献:

[37] 彭岩,张道强.半监督典型相关分析算法.软件学报,2008,19(11):2822−2832. http://www.jos.org.cn/1000-9825/19/2822.htm [doi: 10.3724/SP.J.1001.2008.02822]

陈诗国(1984－),男,河南信阳人,硕士生,主要研究领域为数据挖掘,图像处理.



张道强(1978－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为模式识别,机器学习,数据挖掘.

© 中国科学院软件研究所 http://www.jos.org.cn