半监督类保持局部线性嵌入方法

邓廷权， 王强

引用本文:
邓廷权, 王强. 半监督类保持局部线性嵌入方法[J]. 智能系统学报, 2021, 16(1): 98–107.
DENG Tingquan, WANG Qiang. Semi–supervised class preserving locally linear embedding[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(1): 98–107.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202003007

DOI: 10.11992/tis.202003007

半监督类保持局部线性嵌入方法

邓廷权, 王强

(哈尔滨工程大学 数学科学学院, 黑龙江 哈尔滨 150001)

摘 要: 为使局部线性嵌入(local linear embedding, LLE)这一无监督高维数据的非线性特征提取方法提取出的特征在分类或聚类学习上更优, 提出一种半监督类保持局部线性嵌入 (semi-supervised class preserving local linear embedding, SSCLLE) 的非线性特征提取方法。该方法将半监督信息融入到 LLE 中, 首先对标记样本近邻赋予伪标签, 增大标记样本数量。其次, 对标记样本之间的距离进行局部调整, 缩小同类样本间距, 扩大异类样本间距。同时在局部线性嵌入优化目标函数中增加全局同类样本间距和异类样本间距的约束项, 使得提取出的低维特征可以确保同类样本点互相靠近, 而异类样本点彼此分离。在一系列实验中, 其聚类精确度以及可视化效果明显高于无监督 LLE 和现有半监督流特征提取方法, 表明该方法提取出的特征具有很好的类保持特性。

关键词: 非线性特征提取; 流形学习; 半监督; 标记信息; 聚类; 可视化

中图分类号: TP181 **文献标志码:** A **文章编号:** 1673−4785(2021)01−0098−10

中文引用格式: 邓廷权, 王强. 半监督类保持局部线性嵌入方法 [J]. 智能系统学报, 2021, 16(1): 98–107.

英文引用格式: DENG Tingquan, WANG Qiang. Semi-supervised class preserving locally linear embedding[J]. CAAI transactions on intelligent systems, 2021, 16(1): 98–107.

Semi-supervised class preserving locally linear embedding

DENG Tingquan, WANG Qiang

(College of Mathematical Sciences, Harbin engineering university, Harbin 150001, China)

Abstract: To make local linear embedding (LLE), the nonlinear feature extraction method for unsupervised high-dimensional data, more optimal in classification or clustering learning, we propose a nonlinear semi-supervised class preserving local linear embedding (SSCLLE) feature extraction method. This method integrates semi-supervised information into LLE. First, pseudo-labels are assigned to the nearby neighbors of the labeled samples to increase the number of labeled samples. Second, the distance between the labeled samples is partially adjusted to reduce the distance between similar samples and expand the distance between heterogeneous samples. Simultaneously, the constraints of the globally same sample spacing and heterogeneous sample spacing are added in the local linear embedding optimization objective function so that the extracted low-dimensional features can ensure that the same sample points are near each other, whereas the heterogeneous sample points are separated from each other. In a series of experiments, the clustering accuracy and visualization effect of the proposed method are significantly higher than those of unsupervised LLE and the existing semi-supervised flow feature extraction methods, indicating that the features extracted by this method have good class retention characteristics.

Keywords: nonlinear feature extraction; manifold learning; semi-supervised; labeled information; clustering; visualization

随着信息科技的迅速发展, 数据规模的爆炸式增长成为了大数据时代的主要特征之一。在此时代背景下, 数据通常具有维数高和稀疏性等特点, 为数据挖掘带来了空前的挑战。特征提取作为处理高维数据的有效手段, 通过提取数据的低维特性, 可以将高维特征空间映射到低维特征空间中进行数据的分析和处理, 通常分为线性特征提取和非线性特征提取 2 种方式。非线性特征提取不依赖于线性假设, 对于处理非线性结构的数

收稿日期: 2020−03−04.
基金项目: 国家自然科学基金项目 (11471001, 61872104).
通信作者: 王强. E-mail: 1005834631@qq.com.

据效果较好, 成为当前数据挖掘的热门方向之一。流形学习[1-6]作为一种非线性特征提取方法, 应用了流形在局部结构上与欧氏空间同胚的性质。通过对高维数据样本的分析来挖掘隐藏的本质结构, 从而提取有效的低维特征。然而, 流形学习方法仍然存在一些不足, 例如: 流形学习方法忽略了数据的类别标记信息, 提取的特征并不是分类上的最优特征。因此, 忽略标记信息而提取到的特征在进行数据聚类或分类时, 结果往往与实际存在较大差异。所以希望可以使用半监督[7-14]的方法进行学习, 即少量标记信息来指导特征提取, 同时又使用大量无标记信息的数据点来刻画并保持样本的局部或全局几何、线性等结构。

局部线性嵌入 (LLE)[15]是一种无监督[16]的流形学习方法, 直接用它提取的特征进行数据挖掘如聚类或分类得到的结果并不是很理想。因此我们希望将数据集的标记信息引入到 LLE 方法中用以提高特征提取效果。而已有的一些半监督方法, 例如半监督局部线性嵌入方法 (semi-supervised locally linear embedding, SSLLE) 虽然利用了标记信息对特征提取进行一定的改进, 但它只考虑了近邻点的标记信息做局部调整, 因此当整体标记信息较低时每个近邻中将有可能出现没有标记点的情况。这时 SSLLE 将失去作用并且由于它只考虑近邻的这种调整, 当标记信息很多时它们整体的区分度也不大。本文在 LLE 的基础上利用近邻伪标签赋予得到的标记信息作局部调整, 同时从全局[17]角度对同类数据点和异类数据点进行全局调整, 使得重构数据低维特征空间时, 既保持局部线性结构, 又能使提取后的数据在低维特征空间中可以实现具有相同标记信息的数据点互相靠近, 而标记不同的数据点彼此分离, 从而达到更好的特征提取结果。最后通过聚类分析及可视化证明本文方法的有效性。

1　局部线性嵌入

由 Roweis 等提出的 LLE 是一个经典的保持局部线性特性的流形学习方法, 可以有效提取高维数据的低维特征。其基本原理为: 假设数据是分布在一个流形上的, 任一点均可用它的近邻点经由线性重构而得到。基于局部线性表示系数, 构造优化问题使得数据在高维原始空间到低维特征空间的过程中局部线性重构权值不发生变化, 获得高维数据的低维特征。

假设数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$ 中有 n 个样本点 $x_i,x_i\in\mathbf{R}^m,i\in[1,n]$, $\boldsymbol{Y}\in\mathbf{R}^{n\times d}$ 为特征提取后获得的 n 个低维特征矩阵, $\boldsymbol{Y}=[\boldsymbol{y}_1^{\mathrm{T}}\ \boldsymbol{y}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{y}_i\in\mathbf{R}^d, d\ll m$。

对于每个数据点, 计算每一个数据点 x_i 到其它点的欧氏距离, 找到最近的 k 个点作为该数据样本的近邻, 确定数据的 k 近邻域。也可采用 ε 邻域方法确定数据的近邻点。

假设任一点 x_i 都可用它的 k 近邻通过线性权值 $w_{ij}, j=1,2,\cdots,k$ 加权来得到, 由以下优化问题求解线性重构的权矩阵 $\boldsymbol{w}=[\boldsymbol{w}_1^{\mathrm{T}}\ \boldsymbol{w}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{w}_n^{\mathrm{T}}]^{\mathrm{T}}=(w_{ij})_{k\times n}$, 为

$$\begin{aligned}&\min\varepsilon(\boldsymbol{w})=\sum_{i=1}^{n}\left\|x_i-\sum_{j=1}^{k}w_{ij}x_j\right\|^2\\&\text{s.t.}\quad\sum_{j=1}^{k}w_{ij}=1\end{aligned}\tag{1}$$

容易获得优化问题式 (1) 的最优解:

$$\boldsymbol{w}_i=\frac{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{G}_i^{-1}}{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{G}_i^{-1}\boldsymbol{\Gamma}}\tag{2}$$

式中: $\boldsymbol{G}_i=(\boldsymbol{g}^i{}_{lj})$ 是一个 $k\times k$ 的 Gram 矩阵 (距离矩阵); $\boldsymbol{g}^i{}_{lj}=(x_i-x_{il})(x_i-x_{ij})^{\mathrm{T}}$; $\boldsymbol{\Gamma}=(1,1,\cdots,1)^{\mathrm{T}}$ 是一个 $k\times1$ 的全 1 矩阵; x_{il} 表示样本 x_i 的第 l 个近邻点。记 $\boldsymbol{g}_{il}=x_i-x_{il}$, 则 $\boldsymbol{g}^i{}_{lj}=\boldsymbol{g}_{il}\boldsymbol{g}_{ij}^{\mathrm{T}}$。

基于局部线性重构矩阵式 (2), 构造优化问题:

$$\begin{aligned}&\min\sigma(\boldsymbol{Y})=\sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{y}_j\right\|^2\\&\text{s.t.}\ \sum_{i=1}^{n}\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i=\boldsymbol{I},\quad\sum_{i=1}^{n}\boldsymbol{y}_i=0\end{aligned}\tag{3}$$

获得高维数据 $\boldsymbol{X}$ 的低维嵌入 $\boldsymbol{Y}$。

根据样本的邻域点分布将 k 维行向量 $\boldsymbol{w}_i$ 扩充成 n 维行向量 $\boldsymbol{W}_i$, 记 $\boldsymbol{W}=[\boldsymbol{W}_1^{\mathrm{T}}\ \boldsymbol{W}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{W}_n^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})$, 则优化问题式 (3) 的目标函数可化简为 $\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Y})$。

采用拉格朗日乘子法求解优化问题式 (3), 可得 $\boldsymbol{M}\boldsymbol{Y}=\lambda\boldsymbol{Y}$。即式 (3) 可转化为求特征值问题。实对称半正定矩阵 $\boldsymbol{M}$ 的最小 d 个非 0 特征值对应的特征向量按列排列时, 每行做成的向量的就是对应数据的低维特征 $\boldsymbol{y}_i$。

2　半监督类保持局部线性嵌入方法

在数据挖掘任务中, 监督信息为用户提供强有力的数据分析基础。然而, 众多实际问题只能获得少量样本的监督标记。半监督机器学习方法应运而生。

LLE 是一种经典的无监督高维数据特征提取方法。本文在 LLE 基础上提出一种半监督类保持局部线性嵌入方法 (SSCLLE)。该方法不仅利用近邻伪标签赋予得到的标记信息调整近邻数据间的距离, 而且从全局角度加入了同类数据点和异类数据点的全局约束, 使提取后的数据在低维

特征空间中可以实现具有相同标记信息的数据点互相靠近，而标号不同的数据点彼此分离，达到更好的特征提取效果。

假设 X 是一个半监督数据集，其中少部分数据样本带有标记 (类别标签)。记 X_c 是有标签的数据组成的集合，$l(x)\in\{1,2,\cdots,f\}$ 是 X_c 中各数据点所对应的标签，$L=\{l(x_1),l(x_2),\cdots,l(x_s)\}$，$f$ 是数据集的类数。

一般情况下，X_c 中的样本量较少。在流形学习中，少量监督样本不能全面描述和刻画数据的局部和全局流形结构，致使学习到的特征不能准确反映数据的内在特性。本文给出一种近邻伪标签赋予的方法，给部分未标记样本赋予伪标签，增大标记样本量。

将所有标记样本 X_c 的各自近邻中的未标记点设置与标记点相同的初标签，然后对这些初标签点进行筛选。如果这个未标记点只赋予了一个标签，则将此标签设定为这个点的伪标签。如果这个未标记点有 2 个以上的伪标签，把这个点的所有初标签都去掉，该点依然设定为未标记点，如图 1 所示。

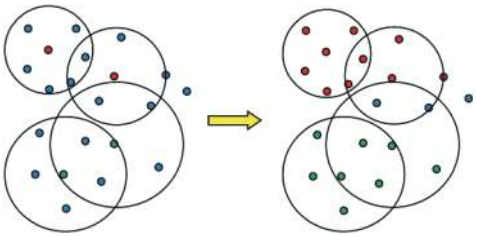


图 1　近邻伪标签赋值方法示意

Fig. 1　Schematic diagram of nearest neighbor pseudo label assignment method

在图 1 的左图中，红色和绿色的点分别代表标记点 (2 类)，蓝色是无标签的点。经过上述近邻伪标签赋值方法后，只有一类标记信息的近邻点保留赋予的标签 (右图新增加的红色点和绿色点)，而有 2 种 (或多种) 标记的近邻点则依旧标为无标记点，保持其蓝色不变 (右图大圆中的 2 个蓝色点)。得到的新标签数据为 X_w，则有标签的数据组成的集合为 $X_z=[X_c,X_w]$，对应的新标签集合为 $L=\{l(x_1),l(x_2),\cdots,l(x_s),l(x_{s+1}),\cdots l(x_{s+t})\}$。

新增加的伪标签虽然不是真实的标签，但由于其与被标注样本具有很好的近邻关系，通过这样的扩充可增加标记信息的量，有利于更好地描述数据的内在结构，发现样本中隐藏的鉴别能力。

为了构造出利用全局信息进行调整的优化问题，首先定义同类数据点对集合：

$$\mathrm{ML}=\{(x_i,x_j)|i\neq j,l(x_i)=l(x_j)\}$$

和异类数据点对集合：

$$\mathrm{CL}=\{(x_i,x_j)|i\neq j,l(x_i)\neq l(x_j)\}$$

分别构造同类样本项偏差和异类样本项偏差：

$$J_{\mathrm{ML}}=\sum_{(x_i,x_j)\in\mathrm{ML}}d^2(\boldsymbol{y}_i,\boldsymbol{y}_j)=\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2$$

$$J_{\mathrm{CL}}=\sum_{(x_i,x_j)\in\mathrm{CL}}d^2(\boldsymbol{y}_i,\boldsymbol{y}_j)=\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2$$

式中 $d(\boldsymbol{y}_i,\boldsymbol{y}_j)$ 表示低维特征 $\boldsymbol{y}_i$ 与 $\boldsymbol{y}_j$ 之间的欧氏距离。

本文的目的是要求同类样本项偏差尽量小，同时确保异类样本项偏差尽可能的大。

构造半监督数据集 X 中每一个数据样本点的线性重构权值。利用数据中已有的标记信息以及新标记的标记信息来重新调整距离矩阵，从而使得构造的数据点的邻域更加有利于提取优质的特征。

$$\boldsymbol{g}_{ij}'=\begin{cases}(1-r)\boldsymbol{g}_{ij}, & l(x_i)=l(x_j)\\(1+r)\boldsymbol{g}_{ij}, & l(x_i)\neq l(x_j)\\\boldsymbol{g}_{ij}, & x_i\text{和}x_j\text{至少一个无标号}\end{cases}\tag{4}$$

式中 $0<r<1$。

从式 (4) 可以看出，如果 2 个样本有相同的类标，则将其距离缩小。如果 2 个样本有不同的类标，则将其距离扩大。在其他情况下，样本点间的距离保持不变。

重置式 (2) 中的距离矩阵为 $\boldsymbol{G}_i'=(\boldsymbol{g}^{i\prime}_{lj})$，其中 $\boldsymbol{g}^{i\prime}_{lj}=\boldsymbol{g}_{il}'\boldsymbol{g}_{ij}'^{\mathrm{T}}$。

再由 (2) 计算样本点的邻域局部线性重构权矩阵由此利用标记信息得到改进后的新重构权矩阵 $\boldsymbol{w}=(w_{ij})$。

基于以上分析，构造如下优化问题：

$$\begin{aligned}\min\rho(\boldsymbol{Y})=&\beta\sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\sum_{j=1}^{n}w_{ij}\boldsymbol{y}_j\right\|^2+\\&\alpha\sum_{(x_i,x_j)\in\mathrm{ML}}d^2(\boldsymbol{y}_i,\boldsymbol{y}_j)-\\&(1-\alpha)\sum_{(x_i,x_j)\in\mathrm{CL}}d^2(\boldsymbol{y}_i,\boldsymbol{y}_j)\\\text{s.t. }&\sum_{i=1}^{n}\boldsymbol{y}_i=0,\ \sum_{i=1}^{n}\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i=\boldsymbol{I}\end{aligned}\tag{5}$$

该优化问题式 (5) 的目标函数由 3 部分组成。第 1 项形式上虽然和 LLE 相同，但其中的重构权矩阵包含了样本点的半监督信息，能够确保提取出的特征既保持数据的局部线性结构不变，又能在局部上使类内 (同类) 数据更紧密，并对类间 (异类) 数据进行分离的效果。第 2 项和第 3 项分别是全局同类样本偏差和全局异类样本偏差，目的是确保同类样本偏差最小，同时确保全局异类样本偏差最大，参数 $\alpha\in(0,1)$ 是 2 个偏差项的平衡系数，权衡同类样本项和异类样本项对目标函数的影响。β 也是一个平衡参数，用于调节局部线性重构对于目标函数的影响。

式 (5) 的约束条件与 LLE 相同，确保提取出

的特征在低维空间中旋转平移伸缩都具有平移和缩放不变性, 其中 $\boldsymbol{I}$ 为 d 阶单位矩阵。

简记式 (5) 的目标函数为

$$\rho(\boldsymbol{Y}) = \beta\sigma(\boldsymbol{Y}) + \alpha\boldsymbol{J}_{\mathrm{ML}} - (1-\alpha)\boldsymbol{J}_{\mathrm{CL}} \tag{6}$$

这样, 式 (6) 的第 1 部分形式上与 LLE 相同, 仍可表示为

$$\sigma(\boldsymbol{Y}) = \sum_{i=1}^{n}\left\|\boldsymbol{y}_i - \sum_{j=1}^{n} W_{ij}\boldsymbol{y}_j\right\|^2 = \mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Y})$$

式中的 $\boldsymbol{M}$ 由式 (2)、(4) 确定。

为了简化第 2 部分和第 3 部分, 给定矩阵[10] $\boldsymbol{Y} = [\boldsymbol{y}_1^{\mathrm{T}}\ \boldsymbol{y}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathbf{R}^{n\times d}$, $\boldsymbol{Z} = [\boldsymbol{z}_1^{\mathrm{T}}\ \boldsymbol{z}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{z}_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathbf{R}^{n\times d}$, 则对任意 $\boldsymbol{y}_i \in \mathbf{R}^d$ 和 $\boldsymbol{z}_j \in \mathbf{R}^d$, 均有:

$$(\boldsymbol{y}_i - \boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{z}_i - \boldsymbol{z}_j) = \sum_{l=1}^{d}(y_{il} - y_{jl})(z_{il} - z_{jl}) = \mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}^{ij}\boldsymbol{Z})$$

其中 $\boldsymbol{A}^{ij} = ((\boldsymbol{A}^{ij})_{pq})$ 为 $n\times n$ 矩阵, 则

$$(\boldsymbol{A}^{ij})_{pq} = \begin{cases} 1, & p = i, q = i \\ 1, & p = j, q = j \\ -1, & p = i, q = j \\ -1, & p = j, q = i \\ 0, & \text{其他} \end{cases}$$

令

$$w_{ij}^{\mathrm{ML}} = \begin{cases} 1, & (x_i, x_j) \in \mathrm{ML} \\ 0, & \text{其他} \end{cases}$$

$$w_{ij}^{\mathrm{CL}} = \begin{cases} 1, & (x_i, x_j) \in \mathrm{CL} \\ 0, & \text{其他} \end{cases}$$

则有:

$$J_{\mathrm{ML}} = \sum\nolimits_{(x_i,x_j)\in\mathrm{ML}} d^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{i,j=1}^{n} w_{ij}^{\mathrm{ML}} d^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{i,j=1}^{n} w_{ij}^{\mathrm{ML}}\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}^{ij}\boldsymbol{Y}) = \mathrm{tr}\left(\boldsymbol{Y}^{\mathrm{T}}\left(\sum_{i,j=1}^{n} w_{ij}^{\mathrm{ML}}\boldsymbol{A}^{ij}\right)\boldsymbol{Y}\right) = \mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{ML}}\boldsymbol{Y})$$

和

$$J_{\mathrm{CL}} = \sum\nolimits_{(x_i,x_j)\in\mathrm{CL}} d^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{i,j=1}^{n} w_{ij}^{\mathrm{CL}} d^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{i,j=1}^{n} w_{ij}^{\mathrm{CL}}\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}^{ij}\boldsymbol{Y}) = \mathrm{tr}\left(Y^{\mathrm{T}}\left(\sum_{i,j=1}^{n} w_{ij}^{\mathrm{CL}}\boldsymbol{A}^{ij}\right)\boldsymbol{Y}\right) = \mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{CL}}\boldsymbol{Y}) \tag{7}$$

因此, 优化问题 (5) 的矩阵表示形式为

$$\min\rho(Y) = \mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{Y})$$
$$\text{s.t. } \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{I},\ \mathbf{1}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{0}$$

式中: $\boldsymbol{H} = \beta\boldsymbol{M} + \alpha\boldsymbol{V}_{\mathrm{ML}} - (1-\alpha)V_{\mathrm{CL}}$; $\mathbf{1} = (1, 1, \cdots, 1)^{\mathrm{T}}$ 是一个 $n\times 1$ 的全 1 矩阵。采用拉格朗日乘子法求解, 优化问题 (7) 的解转化为求解 $\boldsymbol{H}\boldsymbol{Y} = \lambda\boldsymbol{Y}$ 的特征值问题。

计算矩阵 $\boldsymbol{H}$ 的前 d 个最小非零特征值 ($0 \neq \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_d$) 所对应的特征向量 (列向量) $\boldsymbol{v}_p, p = 1, 2, \cdots, d$, 将其构成矩阵 $\boldsymbol{Y} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_p]$, 则矩阵 $\boldsymbol{Y}$ 的第 i 行向量即为高维数据 x_i 的低维特征 y_i。

3 实验及结果分析

为了证明本文提出的 SSCLLE 的性能, 在加州大学欧文分校 (university of california irvine, UCI) 数据集、实物数据集 coil_20 和手写数字 MNIST 数据集上进行实验。实验结果分别与经典的无监督流形学习方法 LLE、半监督 SSLLE[18] 方法, 半监督拉普拉斯特征映射 (semi-supervised laplacian eigenmap, SSLE)[19] 和分类约束降维方法 (classification constrained dimensionality reduction, CCDR)[20] 进行实验对比。从聚类精度和数据可视化角度对它们进行实验比较和分析。

在这里简单介绍 3 种半监督方法。基于 LLE 提出的 SSLLE, 它的思想是结合数据拥有的部分标记信息调整近邻样本点之间的距离, 再利用调整后的距离来重构权值矩阵。虽然 SSLLE 可以利用部分标签信息使得近邻中同类数据点距离更近, 异类数据点更远从而实现更好的分类以及聚类效果。但由于 SSLLE 方法仅对近邻点之间的距离做调整, 缺乏对全局同类异类点的考虑。当标记点较少时近邻中可能出现没有同类或异类的点的情况, 这时 SSLLE 将失去作用。而且由于它只考虑近邻的调整, 当标记信息很多时它们整体的区分度也不大。

SSLE 和 CCDR 都是在拉普拉斯特征映射 (laplacian eigenmap, LE) 的基础上提出的半监督方法。在这里 SSLE 也是一种利用信息在局部做调整的方法, 缺点和 SSLLE 类似。而 CCDR 是一种全局的调整, 相较于 SSLE 有较好的提取效果。

本文 SSCLLE 方法在保持局部线性结构的同时, 不仅利用标记信息对局部做调整, 同时利用全局项对全局做调整。使类内数据更紧密, 而对类间数据进行分离。从而达到更好的特征提取效果, 以下是相关的实验验证。

统一对各方法设定参数, 进行特征提取。这里用聚类精度作为评判方法有效性的指标之一, 利用模糊 C 均值 (fuzzy c-means, FCM) 聚类方法进行聚类分析。关于样本标签个数做以下设置: 从数据集的每类样本中随机抽取 $S(S = 5\%, 10\%, \cdots, 50\%)$ 比例的数据作为已知标签样本。取 20 次实验的平均值作为最终的聚类精度。参数表示: 近邻个数为 k, 低维特征维度为 d, SSLLE 方法调节参数用 r 表示, SSLE方法中的参数用 v 表示, CCDR 方法中的参数用 u 表示, 本文

SSCLLE 方法中 α 和 β 分别用 a 和 b 表示，r 与 SSLLE 中设置相同。

3.1　UCI 中几个数据集

实验中从 UCI 数据库里选 3 个数据集，分别为 Wine 数据集、Seeds 数据集和 WDBC(wisconsin diagnostic breast cancer)。

然后，分别用 5 种方法进行实验比较和分析。根据特征提取的维数 d 做 3 组实验，分别设置 d 的值为 2、3 和 4。每类数据随机标记 5%，每组实验进行 20 次，求聚类精度的平均值来评判 5 种方法的特征提取效果。表 1~3 分别是 d 值为 2、3 和 4 时，各方法对 3 个数据集进行特征提取后得到的平均聚类精度。实验中，将参数设置为：$k=6, r=0.8, v=0.5, u=1, a=0.9, b=10$。

表 1　数据集信息

Table 1　Data set information

数据集	数据个数	属性个数	类别
Wine	178	13	3
Seeds	210	7	3
WDBC	569	30	2

表 2　$d=2$ 时 5 种方法的平均聚类精度

Table 2　Average clustering accuracy of the five methods when d=2　%

数据集	Wine		Seeds		WDBC	
比例/%	5	15	5	15	5	15
LLE	93.44	93.44	76.23	76.23	84.71	84.71
SSLLE	95.17	96.63	**91.191**	91.22	89.09	89.96
SSLE	95.73	97.19	87.703	87.91	89.75	90.9
CCDR	95.62	96.97	88.171	90.01	85.2	90.11
SSCLLE	**96.29**	**97.53**	91.105	**92.19**	**91.53**	**92.09**

表 3　$d=3$ 时 5 种方法的平均聚类精度

Table 3　Average clustering accuracy of the five methods when d=3　%

数据集	Wine		Seeds		WDBC	
比例/%	5	15	5	15	5	15
LLE	94.94	93.94	64.76	64.76	76.77	76.77
SSLLE	94.38	94.49	89.05	89.1	78.03	82.39
SSLE	93.26	93.23	83.81	84.19	63.51	75.04
CCDR	92.92	93.81	86.14	89.05	63.69	79.3
SSCLLE	**95.06**	**95.38**	**89.05**	**90.1**	**78.22**	**86.53**

由表 2~4 数据可知：当特征空间的维数 d 为 3 和 4 时，在 3 个数据集上 SSCLLE 方法的聚类精度都比其他 4 种方法高，其他方法在不同数据集之间聚类精度各有高低。而当 d 为 2 时，虽然 SSCLLE 方法在 Seeds 数据集的实验中的聚类精度并不是全部保持最高，当标记比例为 5% 时 SSLLE 方法仅仅略高于本文方法，在标记比例为 15% 以及另外 2 个数据集时 SSCLLE 的聚类精度最高。总体实验分析可知，本文提出的半监督流形学习方法 SSCLLE 相比无监督方法 LLE 与其他 3 种半监督方法聚类精度最高，体现出本文方法的优势。

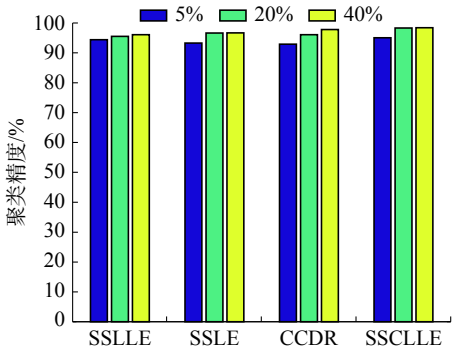
表 4　$d=4$ 时 5 种方法的平均聚类精度

Table 4　Average clustering accuracy of the five methods when d=4　%

数据集	Wine		Seeds		WDBC	
比例/%	5	15	5	15	5	15
LLE	92.13	92.13	72.48	72.48	84.36	84.36
SSLLE	93.38	94.27	83.31	83.52	86.89	86.87
SSLE	87.64	87.75	80.57	80.48	79.37	78.87
CCDR	88.2	90.11	80.43	82.38	75.48	76.94
SSCLLE	**94.18**	**94.27**	**83.81**	**83.76**	**86.96**	**87.99**

对于半监督方法来说标记信息的多少会影响聚类的结果。这里把 3 组 UCI 数据中的每一个类标记信息比例设置为 5%、20% 和 40%，提取特征维数 d=2。图 2 为 3 个数据集在 4 种半监督方法下的实验结果。

由图 2 的实验结果可以看出：3 个数据集的柱状分析图，随着数据的标记比例的增加，各个半监督方法的聚类精度也在增加，符合半监督方法利用越多标记信息就会提高聚类精度的设想。但明显可以看出 2 种基于局部标记信息进行调整的方法 SSLLE 和 SSLE，随着标记信息的增加聚类精度提升，相对考虑全局信息的 SSCLLE 与 CCDR 不明显。而 SSCLLE 方法的聚类精度已经达到了一个很高的值，明显高于 CCDR，所以相对没有 CCDR 提升比率那么高。总体实验分析中可以看到，在每组实验里 SSCLLE 方法的聚类精度基本都能保持最高，证明了本方法在 UCI 数据上的优势。



(a) Wine 数据集

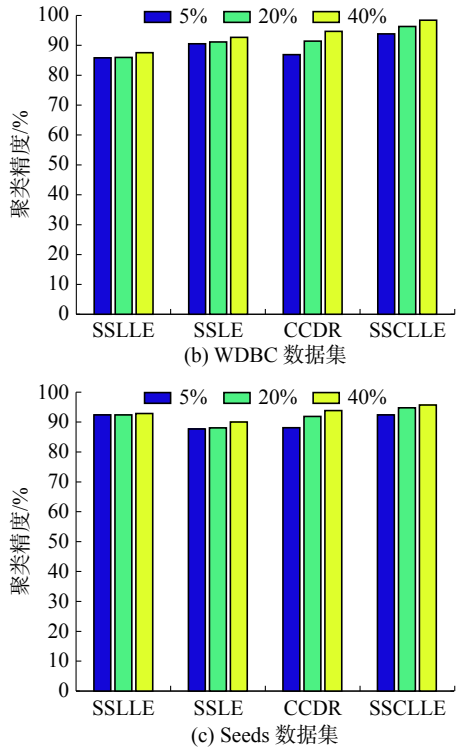


图 2　标记样本的比例对聚类精度的影响，d=2

Fig. 2　Influence of proportion of labeled samples on clustering accuracy, d=2

3.2　实物数据集 COIL _20

这里采用哥伦比亚大学 (COIL-20) 数据集中第 2 种 (背景被丢弃, 图像由包含物体的最小正方形组成), 数据集共有 20 种不同的物体, 每种有 72 张图片。每个图片都是 50×50 的灰度图像, 在实验中将每张图片以行拉成一个 2 500 的向量。最后以向量集的形式进行处理与分析。

从数据集中按顺序选取 6 组数据, 每组 3 类不同的物体。分组分别是 {1,2,3}, {4,5,6}, {7,8,9}, {10,11,12},{13,14,15}和{16,17,18}, 然后再随机选取 3 组不同的数据 {9,7,10}, {7,3,5}, {4,10,1}, 每组运行 20 次计算聚类精度。其中 Group_1~Group_9 分别对应以上 9 组数据, 用不同方法做实验得到聚类精度。参数设置为: $k=8, d=8$, $r=0.5, a=1, b=10, u=1, v=0.5$, 标记比例为 15%, 实验结果如表 5 所示。

由表 5 实验结果可以看到, 在这 9 组数据中由于 SSLLE 和本文方法 SSCLLE 都是在 LLE 方法上进行的一种改进, 所以它们的聚类精度都高于 LLE。且本方法利用了全局标记信息进行调整, 聚类精度明显高于 SSLLE。SSLE 与 CCDR 都是一种在 LE 基础上做的改进, 分析数据可以看出整体上它们略低于 LLE 的改进。且由于 CCDR 也是一种基于全局考虑标记信息的方法, 基本上聚类精度都高于 SSLE。由此体现出基于全局角度考虑标记信息的方法较局部效果要好, 充分说明 SSCLLE 方法基于全局考虑的正确性。除在第 6 组数据中 SSLLE 方法的聚类精度最高外, 其它组中都是本文中提出的 SSCLLE 方法精度最高。

表 5　COIL_20 数据集在不同方法下的平均聚类精度

Table 5　Average clustering accuracy of COIL_20 dataset under different methods　　%

Dataset	LLE	SSLLE	SSLE	CCDR	SSCLLE
Group_1	62.96	63.57	54.17	54.77	**93.52**
Group_2	49.07	51.39	48.19	52.27	**80.56**
Group_3	70.85	71.99	43.56	51.99	**88.89**
Group_4	49.35	52.18	48.94	51.16	**80.09**
Group_5	51.39	51.85	47.22	48.80	**74.54**
Group_6	75.00	**78.03**	44.21	59.17	77.63
Group_7	81.53	86.11	46.71	63.29	**87.50**
Group_8	63.98	63.10	51.02	71.16	**81.02**
Group_9	74.93	75.00	55.93	69.54	**89.81**

接下来随机选出一组数据为{7, 3, 9}, 来做在不同标签比例下不同方法聚类精度的折线图, 参数设置为: $k=7, d=8, a=1, b=10, r=0.5, u=1, v=0.5$。

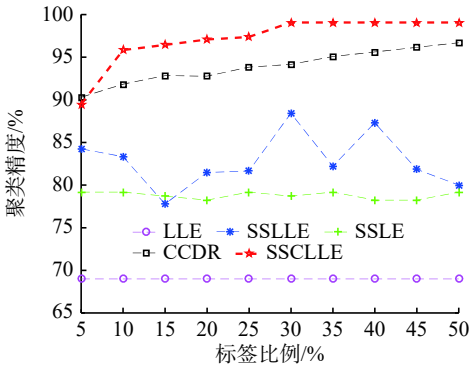


图 3　不同标记比例 COIL_20 数据集聚类精度

Fig. 3　The clustering accuracy of COIL_20 dataset under different labeling ratios

由图 3 可看出在这组数据中随着标记比例的增加无监督 LLE 方法精度保持不变, 而 SSLLE 与 SSLE 方法的聚类精度随着标记比例的增加只发生了波动, 基本没有体现出上升趋势, 说明这 2 种利用类信息只调节近邻关系的方法对一些数据提取到的特征不能很好地提高可分性。而 SSCLLE 和 CCDR 方法都是考虑全局的调整, 可看到聚类精度呈上升趋势, 且高于其他方法, 除在 5% 的情况下略低于 CCDR 方法外, 其余比例下均高于其他方法。体现出 SSCLLE 方法对近邻及全局做调整的优势。

3.3　数据可视化

数据可视化作为一种重要的数据分析方式,

相对于单纯的数据表格等，可更加直观、形象地感知或理解高维数据集的结构分布。为验证SSCLLE方法在可视化上的优势，下面随机选取MNIST数据集中的3个手写数字做可视化实验。分别用LLE方法、半监督：SSLLE、SSLE和CCDR方法，将选取的手写数据集中3个数字提取至2维特征空间中，利用MATLAB画图工具进行画图，同类数据点的颜色和形状一样，分别观察5种不同的方法提取数据点的低维特征分布情况。手写数字选取的是{5,6,8}每类500个点分别将标记比例设为15%，参数设置为：$k=8, d=2, a=1, b=10, r=0.8, u=1, v=0.5$。

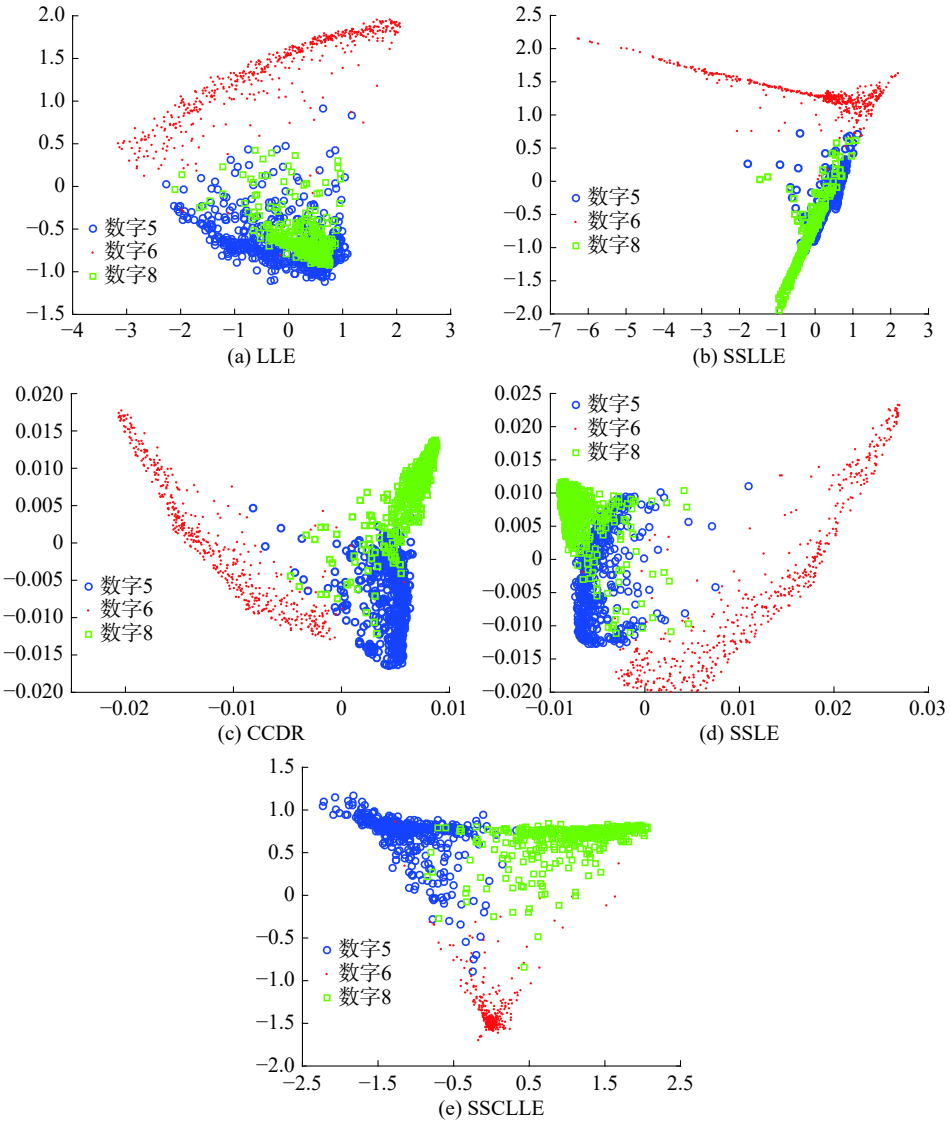


图 4　手写数字可视化

Fig. 4　Visualization of Handwritten digital

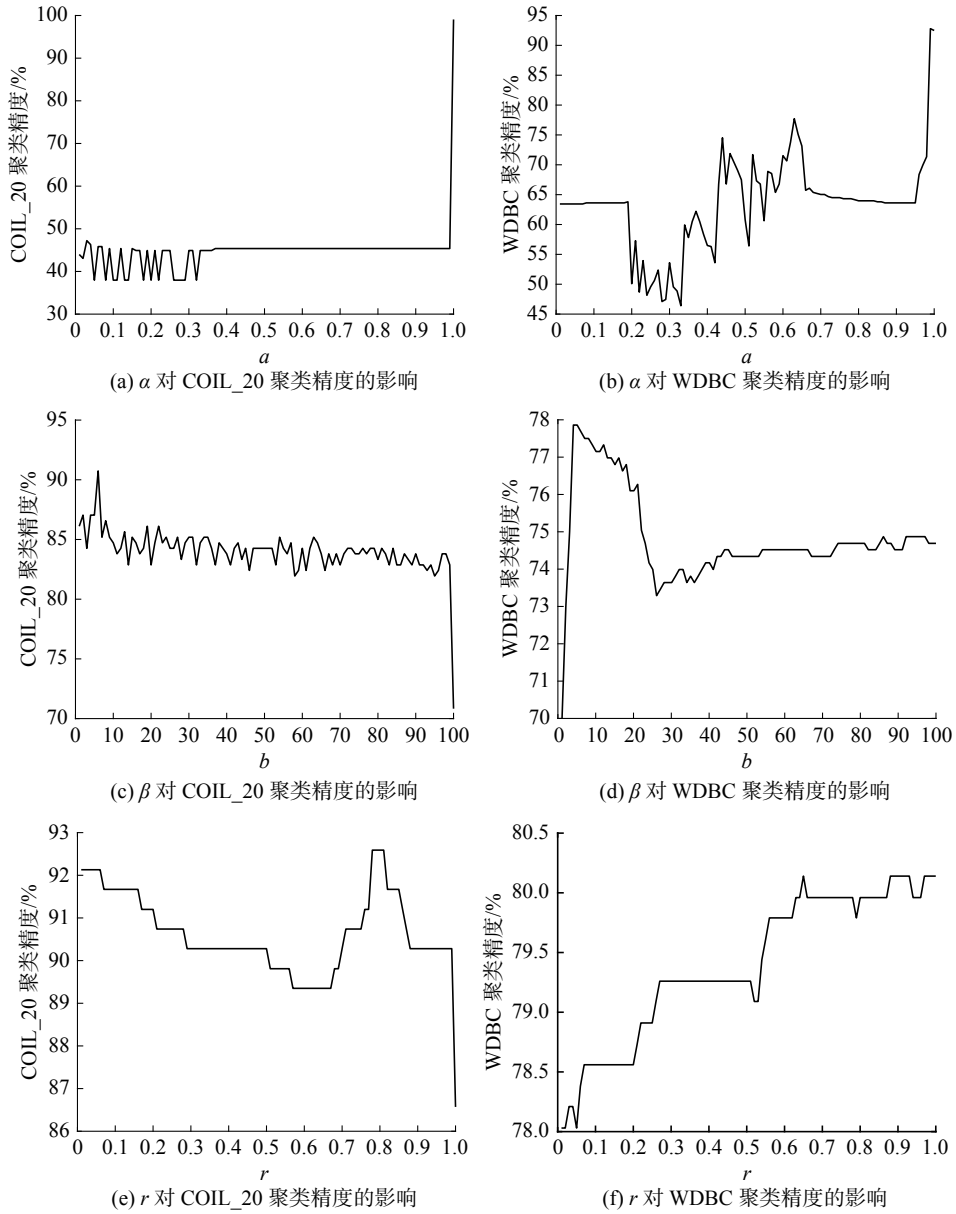
在图4中手写数字的5个可视化图可以看到，无监督的LLE中有2类数据重合部分较大区分度小，因而不利于数据的聚类分析。而基于标记信息局部调整的SSLLE和SSLE的方法相对LLE的分离度明显有所提升，不过依然存在重叠区域。而基于标记信息全局调整的CCDR和本文方法SSCLLE明显3类区分开了，SSCLLE相比CCDR的区分度更高重叠区域最小，可明显区分出3类数据的分布。通过实验可视化的分析，半监督方法在数据可视化方面较无监督方法优势明显，而本文方法的可视化效果相对其他半监督方法效果最好，证明了本文方法的优势。

4　参数影响分析

本方法中参数k、d、α、β和r对特征提取都有影响。k、d参数的选取很多学者都做过讨论，这里不再赘述。本文主要讨论参数α、β和r对特征提取的影响。α和r取[0,1]的实数，α用来

权衡同类样本项和异类样本项对目标函数的影响; β 取大于 0 的值, 用于调节局部线性结构对于目标函数的影响; r 的作用是为了调整标记信息在局部所起到的影响。图 5 展示了随着 α,β 和 r 参数值变化, SSCLLE 方法对于 COIL_20 中的{7,3,9}和 UCI 中 WCBC 数据集特征提取后聚类精度的结果。图 5 中分别用 a、b 表示 α、β。标记比例为 15%, 参数设置为: 在 COIL_20 数据中设定 α=1 ,b = 10 , r=0.8 ; 在 WCBC 数据集中 α=0.99,b = 10 , r=0.7。同时固定其中 2 个参数调整另一个参数, 记录聚类精度的变化。

从图 5 可以看出, 同类数据样本项比异类样本项对聚类精度起到的作用更大。标记比例越高, 异类标记的作用会逐渐增加。在一定的标记比例下, α 一般需要取一个较大的值。在 COIL_20 数据集中当 α 值为 1 时特征提取效果最好, 而在 WDBC 中取值为 0.99 附近时效果最好。β 的取值在 2 个数据集中基本都为 10 时, 得到的聚类精度最高、特征提取效果最好。作为局部调整参数的 r, 相对低于另 2 个参数, 对特征提取的效果也有很大的影响。在 COIL_20 数据集中 r 的取值为 0.8 时效果最好, 在 WDBC 数据集中取 0.9 时效果最好。



(a) α 对 COIL_20 聚类精度的影响　(b) α 对 WDBC 聚类精度的影响

(c) β 对 COIL_20 聚类精度的影响　(d) β 对 WDBC 聚类精度的影响

(e) r 对 COIL_20 聚类精度的影响　(f) r 对 WDBC 聚类精度的影响

图 5　参数 α、β 和 r 对聚类精度的影响

Fig. 5　The influence of parametersα、β and $r$on clustering accuracy

5　t 检验

从手写数字中选取 30 组不同的数据, 每组由 3 个不同的数字组成。对这 30 组数据分别用 5 种方法进行特征提取得到相应的聚类精度。

为了对比 SSCLLE 与其他方法的优劣，利用 SPSS 工具对 SSCLLE 方法得到的聚类精度与其他方法得到的聚类精度做成对 t 检验，得到以下结果如表 6~8 所示。

表 6　配对样本统计

Table 6　Paired sample statistics

配对序号	方法	平均值	个案数	标准差	标准误差平均值
配对 1	LLE	0.899 5	30	0.093 54	0.017 08
	SSCLLE	0.976 0	30	0.012 62	0.002 30
配对 2	SSLLE	0.902 2	30	0.133 21	0.024 32
	SSCLLE	0.976 0	30	0.012 62	0.002 30
配对 3	SSLE	0.907 1	30	0.085 25	0.015 56
	SSCLLE	0.976 0	30	0.012 62	0.002 30
配对 4	CCDR	0.936 8	30	0.086 90	0.015 87
	SSCLLE	0.976 0	30	0.012 62	0.002 30

通过表 8 可以看到 SSCLLE 与其他 4 种方法的显著性均小于 0.05，说明各对比组聚类精度有显著差异。再对比均值，可见本文 SSCLLE 方法相对其他方法能够有效地提高特征提取的效果。

表 7　配对样本相关性

Table 7　Correlation of paired samples

配对序号	方法	个案数	相关性	显著性
配对 1	LLE & SSCLLE	30	0.677	0.000
配对 2	SSLLE & SSCLLE	30	0.542	0.002
配对 3	SSLE & SSCLLE	30	0.703	0.000
配对 4	CCDR & SSCLLE	30	0.667	0.000

6　结束语

本文在 LLE 基础上，提出了一种半监督类保持局部线性嵌入方法 (SSCLLE)。方法中不单考虑了利用近邻伪标签赋予的标记信息对局部近邻做调整，还对样本的全局距离做进一步约束，使其达到既能保持数据的局部线性结构又能使类内数据更紧密，类间数据进行分离，得到很好的特征提取效果。在 UCI 数据集、实物数据集 COIL_20 和手写数据集 MNIST 上对各方法进行实验对比，得到 SSCLLE 方法在聚类精度以及可视化上的结果均高于无监督学习 LLE 方法和半监督学习 SSLLE、SSLE、CCDR 方法。

表 8　配对样本检验

Table 8　Paired sample test

配对序号	方法	配对差值					t	自由度	
		平均值	标准差	标准误差平均值	差值 95% 置信区间				
					下限	上限			
配对 1	LLE-SSCLLE	−0.076 46	0.085 5 0	0.015 6 1	−0.108 3 9	−0.044 54	−4.898	29	0.000
配对 2	SSLLE-SSCLLE	−0.073 85	0.126 8 1	0.023 1 5	−0.121 2 0	−0.026 49	−3.190	29	0.003
配对 3	SSLE-SSCLLE	−0.068 88	0.076 8 9	0.014 0 4	−0.097 5 9	−0.040 16	−4.906	29	0.000
配对 4	CCDR-SSCLLE	−0.039 18	0.079 0 4	0.014 4 3	−0.068 6 9	−0.009 67	−2.715	29	0.011

参考文献：

[1] LIU Feng, ZHANG Weijie, GU Suicheng. Local linear laplacian eigenmaps: a direct extension of LLE[J]. Pattern recognition letters, 2016, 75: 30–35.

[2] JIANG Bo, DING C, LUO Bin. Robust data representation using locally linear embedding guided PCA[J]. Neurocomputing, 2018, 275: 523–532.

[3] WANG Qian, WANG Weiguo, NIAN Rui, et al. Manifold learning in local tangent space via extreme learning machine[J]. Neurocomputing, 2016, 174: 18–30.

[4] TANG Z, LAO H. Robust image hashing via DCT and LLE[J]. Computers and security, 2016, 62: 133–148.

[5] ZHANG Yan, ZHANG Zhao, QIN Jie, et al. Semi-supervised local multi-manifold Isomap by linear embedding for feature extraction[J]. Pattern recognition, 2018, 76: 622–678.

[6] LIU Zhonghua, WANG Xiaohong, PU Jiexin, et a. Non-negative low-rank representation based manifold embedding for semi-supervised learning[J]. Knowledge-based systems, 2017, 136: 121–129.

[7] CHEN Lin, YANG Meng. Semi-supervised dictionary learning with label propagation for image classification[J]. Computational visual media, 2017, 3(1): 83–94.

[8] MIKALSEN K O, SOGUERO-RUIZ C, BIANCHI F M, et al. Noisy multi-label semi-supervised dimensionality

reduction[J]. Pattern recognition, 2019, 90: 257–270.

[9] PARK S H, KIM S B. Active semi-supervised learning with multiple complementary information[J]. Expert systems with applications, 2019, 126: 30–40.

[10] ZHENG Feng, SONG Zhan, SHAO Ling, et al. A semi-supervised approach for dimensionality reduction with distributional similarity[J]. Neurocomputing, 2013, 103: 210–221.

[11] SUN Shiliang, HUSSAIN Z, SHAWE-TAYLOR J. Manifold-preserving graph reduction for sparse semi-supervised learning[J]. Neurocomputing, 2014, 124: 13–21.

[12] KIM K. An improved semi-supervised dimensionality reduction using feature weighting: application to sentiment analysis[J]. Expert systems with applications, 2018, 109: 49–65.

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

[14] LANGLEY P, PAZZANI M J, FISHER D H. Concept formation: knowledge and experience in unsupervised learning[M]. San Mateo: Morgan Kaufmann Publishers, 1991.

[15] YANG Bo, XIANG Ming, ZHANG Yupei. Multi-manifold discriminant isomap for visualization and classification[J]. Pattern recognition, 2016: 215–230.

[16] 张长帅, 周大可, 杨欣. 一种基于核的半监督局部线性嵌入方法 [J]. 计算机工程, 2011, 37(20): 157–159.
ZHANG Changshuai, ZHOU Dake, YANG Xin. Method of kernel-based semi-supervised local linear embedding[J]. Computer engineering, 2011, 37(20): 157–159.

[17] KIM K, LEE J. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction[J]. Pattern recognition, 2014, 47(2): 758–768.

[18] COSTA J A, HERO III A O. Classification constrained dimensionality reduction[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, USA, 2005: 1077–1080.

[19] MARCILLA A, REYES-LABARTA J A, OLAYA M M. Should we trust all the published LLE correlation parameters in phase equilibria? Necessity of their assessment prior to publication[J]. Fluid phase equilibria, 2017, 433: 243–252.

[20] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum, 1981: 35–36.

作者简介:

邓廷权, 教授, 博士生导师, 中国人工智能学会粒计算与知识发现专业委员会委员、黑龙江省工业与应用数学学会副理事长, 主要研究方向为不确定性信息分析理论与方法、机器学习与数据挖掘、模式识别与人工智能。主持和参与国家自然科学基金面上项目各 2 项、主持多项省部级、国家重点实验室基金和横向项目。发表学术论文 100 余篇。

王强, 硕士研究生, 主要研究方向为数据分析理论与方法。