软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):991−1001 [doi: 10.13328/j.cnki.jos.005922]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于选择聚类集成的相似流形学习算法*

罗晓慧, 李凡长, 张 莉, 高家俊

(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

通讯作者: 李凡长, E-mail: lfzh@suda.edu.cn

摘 要: 流形学习是当今最重要的研究方向之一.约简维度的选择影响着流形学习方法的性能.当约简维度恰好是本征维度时,更容易发现原始数据的内在性质.然而,本征维度估计仍然是流形学习的一个研究难点.在此基础上,提出了一种新的无监督方法,即基于选择聚类集成的相似流形学习(SML-SCE)算法,避免了对本征维度的估计,并且性能表现良好.SML-SCE 利用改进的层次平衡 *K*-means(MBKHK)方法生成具有代表性的锚点,高效地构造相似度矩阵.随后计算得到了多个不同维度下的相似低维嵌入,这些低维嵌入是对原始数据的不同表示,而且不同低维嵌入之间的多样性有利于集成学习.因此,SML-SCE 采用选择性聚类集成方法作为结合策略.对于通过 *K*-means 聚类得到的相似低维嵌入的聚类结果,采用聚类间的归一化互信息(NMI)作为权重的衡量标准.最后,舍弃权重较低的聚类,采用基于权重的选择性投票方案,得到最终的聚类结果.在多个数据集的大量实验结果表明了该方法的有效性.

关键词: 相似流形学习;流形学习;集成学习;维度约简

中图法分类号: TP181

中文引用格式: 罗晓慧,李凡长,张莉,高家俊.基于选择聚类集成的相似流形学习算法.软件学报,2020,31(4):991−1001. http://www.jos.org.cn/1000-9825/5922.htm

英文引用格式: Luo XH, Li FZ, Zhang L, Gao JJ. Similar manifold learning based on selective cluster ensemble for image clustering. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):991−1001 (in Chinese). http://www.jos.org.cn/1000-9825/5922.htm

Similar Manifold Learning Based on Selective Cluster Ensemble for Image Clustering

LUO Xiao-Hui, LI Fan-Zhang, ZHANG Li, GAO Jia-Jun

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Manifold learning is one of the most important research directions nowadays. The performance of manifold learning methods is affected by the choice of reduced dimension. When the reduced dimension is the intrinsic dimension, it is easily to handle the original data. However, intrinsic dimension estimation is still a challenge of manifold learning. In this study, a novel unsupervised method is proposed, called similar manifold learning based on selective cluster ensemble (SML-SCE), which avoids the estimation of intrinsic dimension and achieves a promising performance. SML-SCE generates representative anchors with modified balanced K-means based hierarchical *K*-means (MBKHK) to construct similarity matrix efficiently. Moreover, multiple similar low-dimensional embeddings in different dimensions are obtained, which are the different presentations of original data. The diversity of these similar low-dimensional embeddings is benefit to the ensemble learning. Therefore, selective cluster ensemble method is taken advantage of as the combination rule. For the clustering results obtained by *K*-means in similar low-dimensional embeddings, the normalized mutual information (NMI) is

* 基金项目: 国家重点研发计划(2018YFA070170, 2018YFA0701701); 国家自然科学基金(61672364)

Foundation item: National Key Research and Development Program of China (2018YFA070170, 2018YFA0701701); National Natural Science Foundation of China (61672364)

本文由“非经典条件下的机器学习方法”专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-05-29; 修改时间: 2019-08-01; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:19, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0953.007.html

calculated between clusterings as weight. Finally, the low weight clusterings is discarded and a selective vote scheme is adopted based on weight to obtain the final clustering. Extensive experiments on several data sets demonstrate the validity of the proposed method.

Key words: similar manifold learning; manifold learning; ensemble learning; dimensionality reduction

高维数据包含大量冗余信息,直接处理十分耗时.随着数据维数的增加,更会产生“维度灾难”[1]问题.通过维度约简将高维数据映射到低维空间中,不仅可以有效地去除冗余特征[2],而且减少了计算量.目前,维度约简在图像聚类[3]、数据挖掘[4]、机器学习[5,6]等领域发挥着重要作用.

流形学习[7]是一种有效的维度约简方法,近年来流形学习得到迅猛发展,许多经典算法被提了出来,包括等距特征映射(isometric feature mapping,简称 ISOMAP)[8]、局部线性嵌入(local linear embedding,简称 LLE)[9]、拉普拉斯特征映射(Laplacian eigenmaps,简称 LE)[10]、局部保持投影(locality preserving projection,简称 LPP)[11]、谱回归(spectral regression,简称 SR)[12]、无监督大规模图嵌入(unsupervised large graph embedding,简称 ULGE)[13]等等.

流形学习假设原始数据存在嵌入高维空间的低维流形[14],这里,低维流形维度即为本征维度[15].研究发现,约简维度的选择会影响流形学习方法的性能.当约简维度大于本征维度时,得到的低维数据中包含过多的冗余信息;当约简维度小于本征维度时,又会造成数据点在低维空间中重叠,不利于研究其内在性质.如果能够找到高维数据的本征维度,就能轻易地探索数据的内部结构.然而,探究数据的本征维度十分困难.

为了避免本征维度对流形学习方法的影响,本文提出了一种无监督相似流形学习方法,即基于选择聚类集成的相似流形学习(similar manifold learning based on selective cluster ensemble,简称 SML-SCE)算法.SML-SCE 将高维数据映射到不同的低维空间中,获得多个相似流形.不同低维嵌入的多样性更有利于集成学习[16].因此,在 SML-SCE 中采用基于权重的选择聚类集成方法[17]对相似低维流形嵌入的聚类结果进行融合.首先采用 *K*-means 方法对这些相似的低维嵌入进行聚类得到子聚类结果,然后计算子聚类之间的归一化互信息(normalized nutual information,简称 NMI)[18]作为衡量权重的标准.最后,舍弃权重较低的子聚类,采用基于权重的选择性投票方案来获得最终的聚类结果.通过对多个不同维度下相似流形的融合,SML-SCE 避免了对本征维度的估计,同时性能得到提升.

此外,本文还提出了一种新的锚点生成方法——改进的层次平衡 *K*-means(modified balanced *k*-means based hierarchical *K*-means,简称 MBKHK)方法,MBKHK 克服了层次平衡 *K*-means(balanced *K*-means based hierarchical *K*-means,简称 BKHK)[19]方法要求锚点个数必须是 2 的整数次幂的缺点,并选出具有代表性的锚点.

本文的创新点总结如下.

(1) 提出 MBKHK 方法.MBKHK 可生成具有代表性的锚点,并且对锚点个数没有限制.

(2) 提出 SML-SCE 算法.SML-SCE 采用选择聚类集成方法将不同维度下的相似低维流形嵌入的聚类结果进行融合,避免了对本征维度的估计.

本文第 1 节介绍相关工作,包括相似流形的定义、经典流形学习算法,并介绍拉普拉斯特征映射算法.第 2 节详细介绍 SML-SCE 算法,包括锚点生成方法 MBKHK、低维嵌入学习方法以及相似流形结合策略.第 3 节给出 SML-SCE 和其他对比方法的实验结果.第 4 节进行总结和展望.

1 相关工作

1.1 相似流形学习

在介绍相似流形学习之前,首先给出相似流形的定义.

定义 1(相似流形). 设 M 和 N 是两个光滑流形,$f:M\to R^d$ 和 $g:N\to R^d$ 为两个不同的光滑嵌入映射,若对于高维样本集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\},\boldsymbol{x}_i\in R^d$,可由低维空间的样本集合 $Y_M(Y_M\subset M)$和 $Y_N(Y_N\subset N)$分别通过 f 和 g 得到,那么 M 和 N 是相似流形.

在相似流形定义中,f 和 g 表示两个不同的映射,这里的不同可以是由计算方式带来的,也可以是由低维空间

维度不同带来的.而相似流形学习则是在相似流形的基础上,通过对多个低维相似流形进行研究,来探究高维数据的内在性质.

1.2 经典流形学习算法

2000 年,Tenenbaum 在保持流形全局结构的基础上,提出了 ISOMAP 算法,ISOMAP 在计算样本点之间的距离时,采用测地线距离来取代欧氏距离,随后应用经典的多维尺度分析(multidimensional scaling,简称 MDS)[20]算法,保持数据点之间的几何结构不变,最终得到嵌入在高维空间的低维流形.Roweis 和 Saul 考虑保持流形的局部线性结构,提出了局部线性嵌入 LLE 算法,LLE 假设样本点可由局部邻域内的点加权平均重建,并且可以在低维空间内保持样本点的邻域权值不变,从而找出低维流形结构.2001 年,Belkin 和 Niyogi 提出了 LE 算法,LE 利用拉普拉斯-贝尔特拉米算子(Laplacian-Beltrami operator)的特性,求解拉普拉斯 Beltrami 算子的特征函数,得到流形的最优嵌入.2003 年,He 等人在 LE 的基础上,提出了 LPP 算法,LPP 假设高维空间与低维空间之间存在线性投影关系,随后采用与 LE 类似的原理求解,最终得到映射关系,因此,LPP 也被称为 LE 的线性扩展.2007 年,Cai 等人提出了 SR 算法,SR 解决了 LPP 在求解时需要对稠密矩阵进行特征分解的问题,大大减少了计算量.2017 年,Nie 等人提出了 ULGE 算法,ULGE 结合基于锚点策略[21]和无参近邻策略[22],构造了一个对称、双随机、半正定、秩为约简维度的相似度矩阵,随后采用回归方程计算得到投影矩阵,ULGE 降低了计算复杂度且可应用于大规模数据.

1.3 拉普拉斯特征映射(LE)

拉普拉斯特征映射 LE 是一种无监督局部流形学习算法.LE 假设在高维空间距离较近的点映射到低维空间后,仍保持其距离相近,LE 通过保持近邻关系来发现低维流形.对于高维样本点 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}^T\in R^{n\times d}$,$n$ 为样本个数,d 为每个样本点维度.LE 的具体算法步骤如下.

(1) 构造近邻图:LE 通过εNN 或者 kNN 确定近邻点.在εNN 方法中,若$\|\boldsymbol{x}_i-\boldsymbol{x}_j\|\leqslant\varepsilon$,则 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 存在边相连;在 kNN 方法中,若样本点 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的 k 最近邻点,则 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 相连.在此基础上,构造近邻图.

(2) 构造相似度矩阵:常见方法有两种,一种是 0-1 法,即若样本点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间存在边相连,相似度 $a_{ij}=1$,否则为 0;另一种为热核法,即若边存在,相似度 $a_{ij}=\exp\left(-\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_2^2\Big/2\sigma^2\right)$($\sigma$为热核参数),否则,$a_{ij}=0$.

(3) 获得低维嵌入:LE 通过保持近邻关系来获得低维嵌入,其优化的目标函数如下[10]:

$$\min_{\boldsymbol{Y}}\boldsymbol{Y}^T\boldsymbol{L}\boldsymbol{Y}\quad \text{s.t}\ \boldsymbol{Y}^T\boldsymbol{D}\boldsymbol{Y}=\boldsymbol{I} \tag{1}$$

其中,$\boldsymbol{Y}\in R^{n\times p}$ 表示 p 维度下的低维嵌入,p 为约简维度.$\boldsymbol{L}\in R^{n\times n}$ 是拉普拉斯矩阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$.度矩阵 $\boldsymbol{D}\in R^{n\times n}$ 是一个对角矩阵,对角元素定义为 $d_{ii}=\sum_{j=1}^{n}a_{ij}$.公式(1)的最优解 $\boldsymbol{Y}$ 可以转化为广义特征矩阵的特征分解问题.$\boldsymbol{Y}$ 由前 p 个最小非零特征值对应的特征向量组成.

$$\boldsymbol{L}\boldsymbol{Y}=\lambda\boldsymbol{D}\boldsymbol{Y} \tag{2}$$

2 基于选择聚类集成的相似流形学习算法

2.1 锚点生成

传统方法通常直接构造样本点间的相似度矩阵.然而,当数据规模较大时,这种方法的计算复杂度极高.为了解决这一问题,可采用基于锚点的图构造方法,从而降低计算成本.

基于锚点的方法需要生成锚点并计算样本点和锚点之间的相似度矩阵,其中最重要的一步就是锚点的选择.常见方法有随机选择方法和 K-means 方法.随机选择方法计算简单,但不能保证锚点的代表性,性能较差.K-means 方法可以选出具有代表性的锚点,但计算复杂度较高.在数据集规模较大时,性能会受到很大影响.为了解决以上问题,Zhu 等人提出了层次平衡 K-means(BKHK)方法.BKHK 采用平衡二叉树结构,计算复杂度小于 K-means 方法.但是,BHKH 也存在缺点,即 BKHK 要求锚点个数必须是 2 的整数次幂.

本文提出了一种改进的层次平衡 K-means(MBKHK)方法,可避免上述缺点.本质上,MBKHK 是多次运用了

两类平衡 *K*-means 方法.

下面首先介绍两类平衡 *K*-means 方法.在聚类中,平衡 *K*-means 方法要保持簇集内部的点到其中心点的距离尽可能地小,因此得到其目标函数为

$$\min_{\boldsymbol{R}} \sum_{i=1}^{n}\sum_{j=1}^{2} \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|_2^2 r_{ij} \tag{3}$$

其中,$\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2]\in R^{d\times 2}$为中心点矩阵,$\boldsymbol{c}_1$和$\boldsymbol{c}_2$分别为两类的聚类中心点.初始中心点为随机选取.$\boldsymbol{R}\in R^{n\times 2}$为所求的样本点类别矩阵.若 $\boldsymbol{x}_i$ 属于 $\boldsymbol{c}_1$ 类,$r_{i1}=1$ 且 $r_{i2}=0$;若 $\boldsymbol{x}_i$ 属于 $\boldsymbol{c}_2$ 类,$r_{i2}=1$ 且 $r_{i1}=0$.因此,$r_{i1}+r_{i2}=1$.

令α和β分别是两类样本点个数,则$\alpha+\beta=n$.在平衡 *K*-means 方法中,为了能够在两类间迭代地执行平衡 *K*-means 方法,需保持这两类样本点个数基本相等,设 $\alpha=\left\lfloor \frac{n}{2} \right\rfloor$(如果 n 为奇数,则 $\alpha=\frac{n-1}{2}$),$\beta=n-\alpha$.

为了方便起见,定义样本点 $\boldsymbol{X}$ 到中心点 $\boldsymbol{C}$ 的距离为 $\boldsymbol{H}\in R^{n\times 2}$ 且 $h_{ij}=\| \boldsymbol{x}_i - \boldsymbol{c}_j \|_2^2$.因为 $r_{i1}+r_{i2}=1$,所以 r_{i2} 可以被表示为$(1-r_{i1})$.那么,对公式(3)进行化简,可得:

$$\left.\begin{aligned}
\min_{\boldsymbol{R}} \sum_{i=1}^{n}\sum_{j=1}^{2} \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|_2^2 r_{ij} &= \min_{\boldsymbol{R}} \sum_{i=1}^{n} (\left\| \boldsymbol{x}_i - \boldsymbol{c}_1 \right\|_2^2 r_{i1} + \left\| \boldsymbol{x}_i - \boldsymbol{c}_2 \right\|_2^2 r_{i2}) \\
&= \min_{\boldsymbol{R}} \sum_{i=1}^{n} (h_{i1} r_{i1} + h_{i2} r_{i2}) \\
&= \min_{\boldsymbol{R}} \sum_{i=1}^{n} (h_{i1} r_{i1} + h_{i2} (1 - r_{i1})) \\
&= \min_{\boldsymbol{R}} \sum_{i=1}^{n} (r_{i1} (h_{i1} - h_{i2}) + h_{i2}) \\
&\Rightarrow \min_{\boldsymbol{R}} \sum_{i=1}^{n} (r_{i1} (h_{i1} - h_{i2})) \\
&= \min_{\boldsymbol{R}} \boldsymbol{r}_1^T (\boldsymbol{h}_1 - \boldsymbol{h}_2)
\end{aligned}\right\} \tag{4}$$

其中,$\boldsymbol{r}_1$ 表示矩阵 $\boldsymbol{R}$ 的第 1 列.$\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$ 分别为矩阵 $\boldsymbol{H}$ 的第 1 列和第 2 列.

$\boldsymbol{r}_1$ 中仅包含 0 和 1 两个元素,即 $\boldsymbol{r}_1\in\{0,1\}^n$.从公式(4)可以得到:

$$\min_{\boldsymbol{R}} \boldsymbol{r}_1^T (\boldsymbol{h}_1 - \boldsymbol{h}_2) \quad \text{s.t. } \boldsymbol{r}_1 \in \{0,1\}^n, \sum_{i=1}^{n} r_{i1} = \alpha \tag{5}$$

问题(5)的答案很明显.当$(\boldsymbol{h}_1-\boldsymbol{h}_2)$的第 i 个元素是所有元素的前α个最小值时,令 $r_{i1}=1$,除此之外的 r_{i1} 设为 0,$r_{i1}=1-r_{i1}$,此时满足 $\boldsymbol{r}_1^T(\boldsymbol{h}_1-\boldsymbol{h}_2)$ 取得最小值.在得到每个样本点的类别后,计算每一个点到其中心点的距离,取平均值更新两类中心点,重复以上步骤,直到中心点不再变化,此时得到两类的最终聚类结果.

MBKHK 在这两类聚类上分别执行平衡 *K*-means 方法,MBKHK 示意图如图 1 所示.锚点的个数为 m ($m<<n$).下面就 m 的取值分两种情况进行讨论.

(a) m 是 2 的整数次幂 (b) m 不是 2 的整数次幂

Fig.1 Diagram of MBKHK

图 1 MBKHK 示意图

(1) 当 m 是 2 的整数次幂时,层次执行 log$m$ 次平衡 *K*-means 方法,获得 m 个锚点.在这种情况下所产生的锚

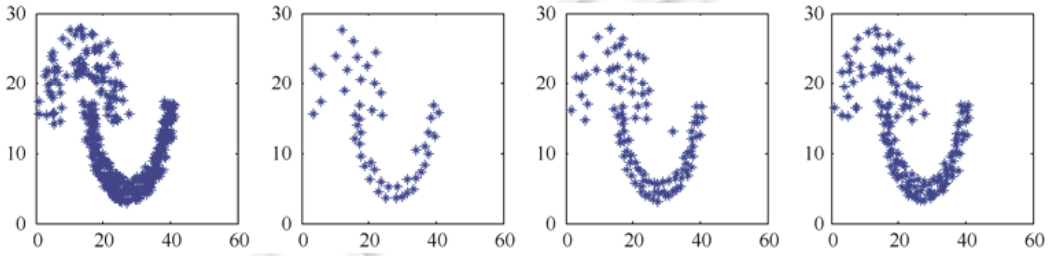
点的示意图如图 1(a)所示.

(2) 当 m 不是 2 的整数次幂时,首先层次执行$\lfloor \log m \rfloor$次平衡 K-means 方法,获得 $2^{\lfloor \log m \rfloor}$个锚点.每一个锚点都是聚类中心点,$\boldsymbol{o}_i$ 即是簇集$\Omega_l(i=1,\ldots,2^{\lfloor \log m \rfloor})$的中心点.每个簇集的样本点到中心点的平均距离可以衡量这个簇集的紧凑性.第 i 个簇集的平均距离 CP_i 表示如下:

$$CP_i = \frac{1}{|\Omega_i|}\sum_{\boldsymbol{x}_j \in \Omega_i} \left\| \boldsymbol{x}_j - \boldsymbol{o}_i \right\| \tag{6}$$

其中,$|\Omega_l|$表示簇集Ω_l的元素个数.CP_i 值越高,说明簇集Ω_l越松散.因此,从 $2^{\lfloor \log m \rfloor}$个簇集中选出前$(m-2^{\lfloor \log m \rfloor})$个 CP_i 值最大的簇集执行平衡 K-means 方法,最终得到 m 个锚点.当 m 不是 2 的整数次幂时,示意图显示如图 1(b)所示.

为了验证 MBKHK 方法的性能,MBKHK 在二维数据集 Jain[23]上进行实验.图 2 显示了 Jain 原始数据和选出的不同数量锚点的分布情况.图 2(a)所示为有 373 个样本点的原始数据情况,图 2(b)~图 2(d)分别为生成的 50、100 和 150 个锚点的分布图.从图中可以明显看出,MBKHK 选出的锚点具有代表性,可以很好地反映原始数据的分布状况.



(a) 原始数据(373) (b) 生成锚点(50) (c) 生成锚点(100) (d) 生成锚点(150)

Fig.2 Jain data set and generated anchors

图 2 Jain 数据集和生成锚点

2.2 低维嵌入学习

获得锚点之后,使用基于锚点的策略构建相似度矩阵.$\boldsymbol{U}\in R^{m\times d}$ 表示锚点矩阵.第 i 个样本点和该样本点的第 j 个最近邻锚点之间的距离用 $d_{ij}=\left\|\boldsymbol{x}_i-\boldsymbol{u}_j\right\|_2^2$ 表示.很明显,有 $d_{i1}\leqslant d_{i2}\leqslant\cdots\leqslant d_{im}$.通过以下公式计算可得到样本点与锚点之间的相似度矩阵 $\boldsymbol{Z}\in R^{n\times m}$[13]:

$$z_{ij}=\begin{cases}\dfrac{d_{i(k+1)}-d_{ij}}{kd_{i(k+1)}-\sum_{j'=1}^{k}d_{ij'}}, & \text{if } \boldsymbol{u}_j\in\Gamma(\boldsymbol{x}_i)\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

其中,$\Gamma(\boldsymbol{x}_i)$是样本点 $\boldsymbol{x}_i$ 的 k 近邻锚点集合.

根据相似度矩阵 $\boldsymbol{Z}$,样本点之间的相似度矩阵 $\boldsymbol{A}\in R^{n\times n}$ 可通过以下公式计算得到[21]:

$$\boldsymbol{A}=\boldsymbol{Z}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}^T \tag{8}$$

其中,$\boldsymbol{\Delta}\in R^{m\times m}$ 是对角矩阵,对角上元素 $\Delta_{jj}=\sum_{i=1}^{n}z_{ij}$.

由文献[21]可知,相似度矩阵 $\boldsymbol{A}$ 是对称、半正定和双随机的.双随机意味着矩阵 $\boldsymbol{A}$ 的各行之和等于 1,各列之和也等于 1,即 $\sum_{i=1}^{n}a_{ij}=\sum_{j=1}^{n}a_{ij}=1$.度矩阵 $\boldsymbol{D}$ 对角上元素定义为 $d_{ii}=\sum_{j=1}^{n}a_{ij}$.根据 $\sum_{j=1}^{n}a_{ij}=1$,度矩阵的对角元素值等于 1,此时,$\boldsymbol{D}=\boldsymbol{I}$,$\boldsymbol{I}\in R^{n\times n}$ 为单位矩阵.$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{A}$,$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{A}$ 将带入 LE 的目标函数(1),可得:

$$\min_{\boldsymbol{Y}}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{Y}\quad \text{s.t } \boldsymbol{Y}^T\boldsymbol{I}\boldsymbol{Y}=\boldsymbol{I} \tag{9}$$

公式(9)可以化简为

$$\max_{\boldsymbol{Y}} \boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \quad \text{s.t } \boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{I} \tag{10}$$

公式(10)的最优解 $\boldsymbol{Y}$ 可以转化为特征矩阵 $\boldsymbol{AY}=\lambda \boldsymbol{Y}$ 的特征分解问题.$\boldsymbol{Y}$ 由矩阵 $\boldsymbol{A}$ 的前 p 个最大特征值对应的特征向量组成.

重复以上步骤,可以获得不同维度下的低维流形嵌入,维度范围记为 $[\kappa, t]$.

2.3 相似流形结合

对于高维数据,可以通过约简维度进行研究.在本征维度下,数据的内部性质更容易被探究.然而,本征维度的研究十分耗时.SML-SCE 通过将高维数据映射到不同低维空间,然后采用选择聚类集成方法将这些相似的低维嵌入进行结合.SML-SCE 避免了对本征维度的寻找,且取得了良好的效果.

采用 *K*-means 方法对 t 个相似低维嵌入进行聚类,$t=t-\kappa+1$,获得多个聚类结果.但是,*K*-means 方法得到的聚类结果是混乱的.例如,对于聚类标签向量$[1,1,1,2,2,3,]^T$ 和$[2,2,2,3,3,1]^T$,两者有着不同的呈现,但却表达了同一种聚类结果.为了结合这些聚类结果,首先需要对聚类标签向量进行校正.一般而言,有较强对应关系的聚类标记覆盖相同对象的个数应该是最大的.例如,标签向量$[1,1,1,2,2,3,]^T$ 中的标记{1}和标签向量$[2,2,2,3,3,1]^T$ 中的标记{2}.将第 2 个聚类标签向量向第 1 个聚类标签向量匹配,找到具有最大相同覆盖对象的对应标记,使得第 2 个聚类标签向量中的该标记等于第 1 个聚类标签向量中的对应标记,并记入新标签向量中.随后,在两个标签向量中分别移除已匹配的标记.重复以上过程,直到第 2 个聚类标签向量中的所有标记均在第 1 个聚类标签向量中找到对应的标记.该校正方法被称为 bestMap 方法.bestMap 方法举例见表 1.

Table 1 Example of bestMap method

表 1 bestMap 方法举例

	标签向量 1	标签向量 2	最大相同覆盖对象个数	标签向量 1 对应标记	标签向量 2 对应标记	新标签向量
第 1 轮	$[1,1,1,2,2,3]^T$	$[2,2,2,3,3,1]^T$	3	{1}	{2}	$[1,1,1,\varnothing,\varnothing,\varnothing]^T$
第 2 轮	$[\varnothing,\varnothing,\varnothing,2,2,3]^T$	$[\varnothing,\varnothing,\varnothing,3,3,1]^T$	2	{2}	{3}	$[1,1,1,2,2,\varnothing]^T$
第 3 轮	$[\varnothing,\varnothing,\varnothing,\varnothing,\varnothing,3]^T$	$[\varnothing,\varnothing,\varnothing,\varnothing,\varnothing,1]^T$	1	{3}	{1}	$[1,1,1,2,2,3]^T$

对于多个聚类标签向量,从中随机选择一个作为参考标签向量,并将剩下的标签向量与参考标签向量相匹配,得到校正后的聚类标签向量.

随后,采用基于权重的选择聚类集成方法来结合这些聚类结果.聚类标签向量之间的归一化互信息 NMI 在一定程度上可以描述聚类之间的紧密性.因此,在本算法中,NMI 被用作聚类间衡量权重的标准.

由 *K*-means 方法得到的聚类标签向量 $\boldsymbol{\eta}_a \in \mathbb{R}^n$ 和 $\boldsymbol{\eta}_b \in \mathbb{R}^n$, 其聚类标记分别为 $\{C_a^1, C_a^2, ..., C_a^K\}$ 和 $\{C_b^1, C_b^2, ..., C_b^K\}$. 假设聚类标记 C_a^i 和 C_b^j 分别包含 n_i 和 n_j 个元素,那么很明显 $\sum_{i=1}^K n_i = \sum_{j=1}^K n_j = n$. 将在 C_a^i 和 C_b^j 中共同含有的元素个数记为 n_{ij}.两类之间的 NMI 定义为

$$\Phi^{NMI}(\boldsymbol{\eta}_a, \boldsymbol{\eta}_b) = \frac{2}{n} \sum_{i=1}^K \sum_{j=1}^K n_{ij} \log_{K^2} \frac{n_{ij} n}{n_i n_j} \tag{11}$$

聚类标签向量 $\boldsymbol{\eta}_i (i = 1,2,...,t)$ 的平均 NMI 可通过公式(12)计算得到:

$$\gamma_i = \frac{1}{t-1} \sum_{j=1, j \neq i}^{t} \Phi^{NMI}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j) \tag{12}$$

γ_i 的值越高,聚类 $\boldsymbol{\eta}_i$ 中包含的信息越多.因此,$\boldsymbol{\eta}_i$ 的权重可通过以下公式计算得到:

$$w_i = \frac{\gamma_i}{\sum_{j=1}^t \gamma_j} \tag{13}$$

很明显,各聚类的权重之和等于 1,即 $\sum_{i=1}^t w_i = 1$.

当聚类结果的权重过低时,则意味着这个聚类结果对集成学习有负面影响,因此需要将其舍弃.换言之,将该聚类的权重置为 0.在本算法中,按照百分比来选择权重,舍弃率标记为 *dr*.最终,采用基于权重的选择性投票方案,得到最终的聚类结果.

$$FC(\boldsymbol{x}_i)=\arg\max_{q\in\{1,2,\ldots,K\}}\sum_{j=1}^{t}w_j I(\eta_j^i=q) \tag{14}$$

其中,K 是样本数据集 $\boldsymbol{X}$ 的类别数,η_j^i 表示聚类 $\boldsymbol{\eta}_j$ 的第 i 个元素.$I(\cdot)$是一个指示矩阵,当“$\cdot$”为真时,$I(\cdot)$=1,反之,当“$\cdot$”为假时,$I(\cdot)$等于 0.

SML-SCE 方法的具体流程见算法 1.

算法 1. 基于选择聚类集成的相似流形学习算法(SML-SCE).

输入:样本数据集 $\boldsymbol{X}\in R^{n\times d}$,低维嵌入维度范围$[\kappa,\iota]$,锚点个数 m,权重舍弃率 dr.

输出:最终聚类结果.

1. 获得 $t=\iota-\kappa+1$ 个相似低维流形嵌入子聚类:

   For $i=\kappa,\ldots,\iota$ **Do**

   (a) 执行 MBKHK,获得 m 个锚点;

   (b) 根据公式(7)和公式(8)计算相似度矩阵 $\boldsymbol{A}$;

   (c) 对矩阵 $\boldsymbol{A}$ 进行特征分解,得到 i 维度下的低维嵌入 $\boldsymbol{Y}_i$;

   (d) 对低维嵌入 $\boldsymbol{Y}_i$ 进行 K-means 聚类,得到子聚类结果;

   End For

2. 采用 bestMap 方法对 t 个聚类结果进行校正;
3. 根据公式(12)和公式(13)计算聚类之间的 NMI 并得到权重;
4. 将聚类中权重较低的 dr 舍弃,这些舍弃聚类的权重置为 0;
5. 根据公式(14)采用基于权重的选择投票策略来获得最终聚类结果.

3 实验结果与分析

为了验证 SML-SCE 算法的性能,将其与多种方法进行对比,包括主成分分析(principal component analysis,简称 PCA)[24]、LPP、SR、ULGE、图嵌入集成学习(graph embedding-based ensemble learning,简称 GEEL)[25]算法、基于图嵌入的无监督集成学习(unsupervised ensemble learning based on graph ebedding,简称 UEL-GE)[26]算法.其中,PCA、LPP、SR、ULGE 是流形学习算法,GEEL 和 UEL-GE 是集成学习算法.此外,K-means 方法直接在高维数据进行聚类,作为一种 Baseline 方法.

3.1 数据集介绍

实验分别在 3 个图像数据集上进行,分别为 DBRHD[27]、COIL20[28]和 ETH[29].DBRHD 是手写体数字数据集,包含 0～9 这 10 个数字,一共 10 个类别.COIL20 是哥伦比亚大学图像库数据集,包含 20 个物品,每个物品拥有 72 张不同角度的图片.ETH 是物品聚类和识别数据集,包含 80 个物品,一共归属为 8 个类别.这 3 个数据集的详细信息见表 2.

Table 2 Description of data sets

表 2 数据集细节描述

数据集	样本点数	数据维度	类别个数
DBRHD	10 992	16	10
COIL20	1 440	1 024	20
ETH	3 280	1 024	8

3.2 参数设置

为了验证 SML-SCE 和对比方法的性能,在不同的低维空间进行实验.SML-SCE 结合多个相似低维嵌入,令低维嵌入的维度范围为 1~15,即,κ=1,ι=15.在构建图时,LPP、SR 中的近邻样本点数和 ULGE、GEEL、UEL-GE、SML-SCE 中的近邻锚点数相同,都等于 5[13].LPP 和 SR 中的高斯核参数设置为 1.在 SR、ULGE 和 UEL-GE 的正则化参数等于 0.01[13].GEEL 和 UEL-GE 中的个体学习器的个数为 7[25].ULGE、GEEL、UEL-GE 和 SML-SCE

都有关于锚点个数的参数.在 4 种方法中,锚点个数都等于数据集样本点个数的 35%.在 SML-SCE 中,权重的丢弃率 *dr* 设置为 70%.

3.3 评价指标

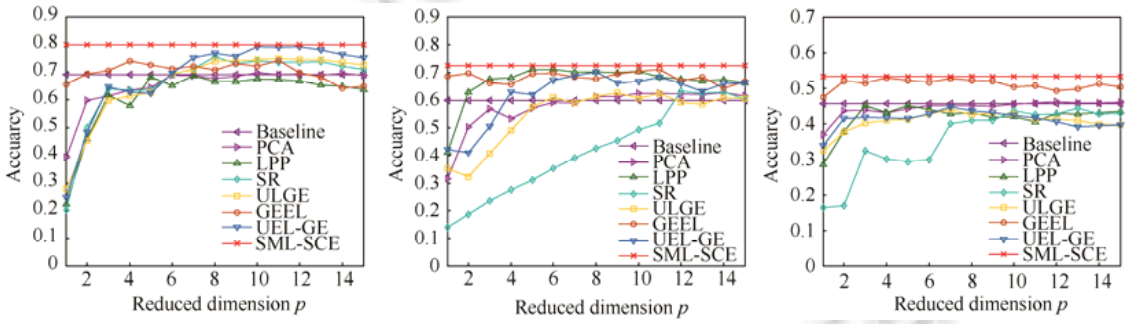
为了比较各种算法之间的性能,采用 *K*-means 方法对低维嵌入进行聚类.聚类结果可通过聚类准确率(accuracy,简称 ACC)和归一化互信息(NMI)来衡量,ACC 和 NMI 的值越高,表示聚类效果越好.将聚类结果和真实标签进行比较,分别计算 ACC 和 NMI.为了保证实验结果的准确性,每种方法运行 20 次,并记录平均值.

所有实验代码均采用 MATLAB 编写.实验硬件环境为 3.19GHz,Intel(R) Core(TM) i5-6500 CPU,8GB 内存,系统为 Windows 10.

3.4 结果分析

图 3 是在不同约简维度下所有方法的 ACC 曲线.图 3(a)~图 3(c)分别为在 DBRHD、COIL20 和 ETH 数据集上的实验结果.因为 SML-SCE 结合了从 1 维到 15 维的相似低维嵌入,所以在图中显示为一条直线.从图中可以发现,SML-SCE 在 3 个数据集上的性能均优于所有对比方法.无论降到多少维度,SML-SCE 均表现出明显的优越性.

DBRHD 数据集共含有 10 类,从中随机选取 2/4/6/8/10 类作为新数据集进行实验,实验结果呈现在表 3 中.相似地,从 COIL20 中随机选取 5/10/15/20 类,从 ETH 中随机选取 2/4/6/8 类作为新数据集进行实验,将 SML-SCE 与其他方法进行对比.在实验中,对比算法的约简维度设置为数据集类别数.表 3~表 5 分别记录了在 DBRHD、COIL20 和 ETH 数据集上的平均 ACC 和 NMI.在同一数据集下表现最好方法的实验结果加粗显示.由表格可知,在 3 个数据集的不同类别下,SML-SCE 均表现最优,这表明 SML-SCE 算法明显优于其他对比算法.



(a) DBRHD (b) COIL20 (c) ETH

Fig.3 Accuracy under different reduced dimensions *p* on three data sets

图 3 3 个数据集不同维度下的聚类准确率 ACC

Table 3 The performance of different classes on DBRHD

表 3 DBRHD 数据集上不同类别下的聚类结果

方法	ACC (%)					NMI (%)				
	2 类	4 类	6 类	8 类	10 类	2 类	4 类	6 类	8 类	10 类
Baseline	90.98	75.99	70.41	69.31	69.37	67.35	65.19	63.02	66.60	68.20
PCA	91.42	76.76	69.21	69.88	69.95	68.26	65.40	62.25	66.85	68.07
LPP	74.72	72.95	66.27	64.67	67.54	34.54	62.27	60.48	59.52	66.26
SR	89.40	76.83	72.55	74.32	74.09	66.34	69.96	66.65	71.46	71.19
ULGE	81.30	75.55	71.76	75.07	74.68	47.38	65.19	65.95	71.91	71.73
GEEL	57.46	56.29	62.75	70.83	72.30	8.30	43.25	63.10	73.17	77.49
UEL-GE	81.74	77.53	72.14	74.77	79.26	46.86	67.26	65.82	70.74	73.02
SML-SCE	**92.64**	**78.16**	**78.22**	**78.31**	**80.05**	**73.28**	**74.96**	**76.90**	**79.89**	**81.70**

Table 4 The performance of different classes on COIL20

表 4 COIL20 数据集上不同类别下的聚类结果

方法	ACC (%)				NMI (%)			
	5 类	10 类	15 类	20 类	5 类	10 类	15 类	20 类
Baseline	70.53	63.14	63.53	60.29	72.63	74.08	77.07	76.67
PCA	71.72	63.82	65.18	61.50	73.32	74.13	77.41	76.52
LPP	71.11	66.50	67.01	63.59	74.12	78.60	80.94	79.74
SR	66.40	59.08	62.44	59.77	77.29	80.00	85.56	85.17
ULGE	42.63	43.80	51.71	54.66	35.83	58.23	71.45	76.93
GEEL	29.53	54.01	55.22	61.34	18.20	65.96	74.89	80.98
UEL-GE	29.64	48.35	54.33	66.53	15.75	58.41	70.26	82.78
SML-SCE	**78.44**	**74.31**	**76.32**	**71.64**	**84.36**	**84.55**	**87.72**	**85.54**

Table 5 The performance of different classes on ETH

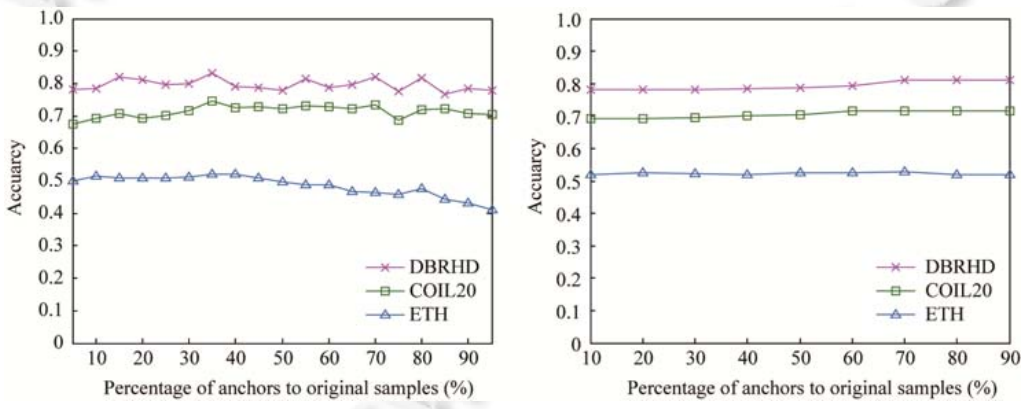
表 5 ETH 数据集上不同类别下的聚类结果

方法	ACC (%)				NMI (%)			
	2 类	4 类	6 类	8 类	2 类	4 类	6 类	8 类
Baseline	74.53	63.74	54.04	45.70	39.97	48.12	47.52	44.01
PCA	74.49	63.72	53.61	45.16	39.93	47.67	46.80	43.02
LPP	60.31	46.65	33.59	43.15	5.04	18.06	16.43	44.27
SR	70.32	55.86	46.73	41.08	33.35	44.55	45.74	42.93
ULGE	60.81	55.21	50.98	42.55	11.53	41.60	49.24	43.93
GEEL	61.26	54.74	57.11	52.28	12.91	41.28	57.35	54.51
UEL-GE	56.66	60.04	51.69	43.79	7.01	46.43	49.89	44.19
SML-SCE	**75.76**	**68.01**	**58.30**	**53.30**	**44.66**	**64.85**	**61.45**	**56.25**

3.5 参数分析

这一节讨论对算法性能产生影响的两个参数,即锚点个数 m 和权重的舍弃率 dr.图 4(a)展示了 3 个数据集上不同锚点个数下的 ACC 曲线.设锚点取值范围为样本点总数的 10%~90%.从图中可以观察到,当 m 取值为 35%时,在 3 个数据集上 ACC 均取得最大值.并且,大量的锚点对于最终性能来说是无用的,甚至会使得 ACC 值有所下降.因此,在 SML-SCE 中,锚点取值为样本点总数的 35%.

图 4(b)是在 3 个数据集上不同权重舍弃率 dr 下的 ACC 曲线.观察图像可知,在不同舍弃率 dr 下 ACC 的变化并不明显,但是,当舍弃率取值为 70%时,SML-SCE 在 3 个数据集上均表现最好.基于此,实验中舍弃率 dr 取值为 70%.



(a) 不同锚点 m 下的聚类准确率 ACC (b) 不同舍弃率 m 下的聚类准确率 ACC

Fig.4 Accuracy under different parameters on three data sets

图 4 3 个数据集不同参数下的聚类准确率 ACC

4 总结与展望

本文提出了一种新型的锚点生成方法,名为改进的层次平衡 *K*-means(MBKHK)方法.MBKHK 生成了具有代表性的锚点,且对锚点的个数没有限制.此外,文中提出了相似流形的概念,并提出基于选择聚类集成的相似流形学习(SML-SCE)算法.SML-SCE 利用 MBKHK 生成锚点,构造相似度矩阵,直接得到低维嵌入,并采用选择聚类集成融合多个不同维度的相似低维流形,得到最终聚类结果.SML-SCE 避免了对约简维度的选择和对本征维度的估计,并取得了优异的性能.在 3 个图像数据集上的实验结果也表明了 SML-SCE 算法的优越性.

SML-SCE 是一种无监督的相似流形算法,但在现实生活中,大多数数据并不全都是无标签的未知数据,因此,未来可以结合部分有标签数据和大多数无标签数据,对本文算法进行改进,将其拓展到半监督领域.

References:

[1] Keogh E, Mueen A. Curse of dimensionality. In: Encyclopedia of Machine Learning and Data Mining. 2017. 314–315.

[2] Choi JY, Bae SH, Qiu X, Fox G. High performance dimension reduction and visualization for large high-dimensional data analysis. In: Proc. of the IEEE/ACM Int'l Conf. on Cluster, Cloud and Grid Computing. 2010. 331–340.

[3] Liu H, Ming S, Sheng L. Infinite ensemble for image clustering. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 1745–1754.

[4] Lu H, Setiono R, Liu H. Effective data mining using neural networks. IEEE Trans. on Knowledge and Data Engineering, 1996,8(6): 957–961.

[5] Singh A, Ganapathysubramanian B, Singh AK. Machine learning for high-throughput stress phenotyping in plants. Trends in Plant Science, 2016,21(2):110–124.

[6] Li FZ, Zhang L, Zhang Z. Lie Group Machine Learning. Walter de Gruyter GmbH and Co KG, 2018.

[7] Zhao Y, You X, Yu S. Multi-view manifold learning with locality alignment. Pattern Recognition, 2018,78:154–166.

[8] Tenenbaum JB, Silva VD, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290 (5500):2319–2323.

[9] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding, Science, 2000,290(5500):2323–2326.

[10] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. of the Int'l Conf. on Neural Information Processing Systems: Natural and Synthetic. 2002. 585–591.

[11] He XF, Niyogi P. Locality preserving projections. Advances in Neural Information Processing Systems, 2003,16(1):186–197.

[12] Cai D, He XF, Han J. Spectral regression: A unified subspace learning framework for content-based image retrieval. In: Proc. of the ACM Int'l Conf. on Multimedia. 2007,60:403–412.

[13] Nie FP, Zhu W, Li XL. Unsupervised large graph embedding. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 2422–2428.

[14] Li YY. Curvature-aware manifold learning. Pattern Recognition, 2018,83:273–286.

[15] Costa JA, Hero AO. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In: Proc. of the European Signal Processing Conf. 2004. 369–372.

[16] Zhou ZH. When semi-supervised learning meets ensemble learning. Frontiers of Electrical and Electronic Engineering in China, 2011,6(1):6–16.

[17] Tang W, Zhou ZH. Bagging-based selective clusterer ensemble. Ruan Jian Xue Bao/Journal of Software, 2005,16(4):496–502 (in Chinese with English abstract). [doi: 10.1360/jos160496]

[18] Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans. on Neural Networks, 2009, 20(2):189–201.

[19] Zhu W, Nie FP, Li X. Fast spectral clustering with efficient large graph construction. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 2017. 2492–2496.

[20] Cox TF, Cox MA. Multidimensional Scaling. Chapman and Hall, 2000.

[21] Liu W, He J, Chang SF. Large graph construction for scalable semi-supervised learning. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 679–686.

[22] Nie FP, Wang X, Jordan MI, Huang H. The constrained Laplacian rank algorithm for graph-based clustering. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016. 1969–1976.

[23] Jain AK, Law MHC. Data clustering: A user's Dilemma. In: Proc. of the Int'l Conf. on Pattern Recognition and Machine Intelligence. 2005,3776:1–10.

[24] Jolliffe IT. Principal component analysis. Journal of Marketing Research, 2002,87(100):513.

[25] Luo XH, Zhang L, Li FZ, Wang BJ. Graph embedding-based ensemble learning for image clustering. In: Proc. of the 24th Int'l Conf. on Pattern Recognition. 2018. 213–218.

[26] Luo XH, Zhang L, Li FZ, Hu CX. Unsupervised ensemble learning based on graph embedding for image clustering. In: Proc. of the Int'l Conf. on Neural Information Processing. 2018,11303:38–47.

[27] Alimoglu F, Alpaydin E, Denizhan Y. Combining multiple classifiers for pen-based handwritten digit recognition. In: Proc. of the 4th Int'l Conf. on Document Analysis and Recognition. 1996,2:637–640.

[28] Nene SA, Nayar SK, Murase H. Columbia object image library (coil-20). Columbia University, 1996. http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[29] Leibe B, Schiele B. Interleaving object categorization and segmentation. Cognitive Vision Systems, 2006,3948:145–161.

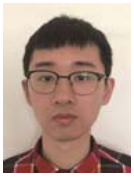
附中文参考文献:

[17] 唐伟,周志华.基于 Bagging 的选择性聚类集成.软件学报,2005,16(4):496–502. [doi: 10.1360/jos160496]

罗晓慧(1994－),女,江苏省泰州人,硕士,CCF 学生会员,主要研究领域为流形学习.

张莉(1975－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,模式识别.

李凡长(1964－),男,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能.



高家俊(1995－),男,硕士,CCF 学生会员,主要研究领域为流形学习.