Journal of Computer Applications
计算机应用,2021,41(3):686－693
ISSN 1001-9081
CODEN JYIIDU
2021-03-10
http://www.joca.cn

文章编号:1001-9081(2021)03-0686-08
DOI:10.11772/j.issn.1001-9081.2020071095

结合改进密度峰值聚类和共享子空间的协同训练算法

吕　佳[1,2*],鲜　焱[1,2]
(1.重庆师范大学 计算机与信息科学学院,重庆 401331;　2.重庆师范大学 重庆市数字农业服务工程技术研究中心,重庆 401331)
(*通信作者电子邮箱lvjia@cqnu.edu.cn)

摘　要:针对协同训练算法在迭代过程中加入的无标记样本的有用信息不足和多分类器对样本标记不一致导致的分类错误累积问题,提出结合改进密度峰值聚类和共享子空间的协同训练算法。该算法先采取属性集合互补的方式得到两个基分类器,然后基于虹吸平衡法则进行改进密度峰值聚类,并从簇中心出发来推进式选择相互邻近度高的无标记样本交由两个基分类器进行分类,最后利用多视图非负矩阵分解算法得到的共享子空间来确定标记不一致样本的最终类别。该算法利用改进密度峰值聚类和相互邻近度选择出更具空间结构代表性的无标记样本,并采用共享子空间来修订标记不一致的样本,解决了因样本误分类造成的分类精度低的问题。在9个UCI数据集上的多组对比实验证明了该算法的有效性,实验结果表明所提算法相较于对比算法在7个数据集上取得最高的分类正确率,在另2个数据集取得次高的分类正确率。

关键词:协同训练;密度峰值聚类;虹吸平衡法则;共享子空间;相互邻近度
中图分类号:TP181　　**文献标志码:**A

Co-training algorithm combining improved density peak clustering and shared subspace

LYU Jia[1,2*], XIAN Yan[1,2]
(1. *College of Computer and Information Science, Chongqing Normal University, Chongqing* 401331, *China*;
2. *Chongqing Center of Engineering Technology Research on Digital Agriculture Service, Chongqing Normal University, Chongqing* 401331, *China*)

Abstract: There would be lack of useful information in added unlabeled samples during the iterations of co-training algorithm, meanwhile, the labels of the samples labeled by multiple classifiers may happen to be inconsistent, which would lead to accumulation of classification errors. To solve the above problems, a co-training algorithm combining improved density peak clustering and shared subspace was proposed. Firstly, the two base classifiers were obtained by the complementation of attribute sets. Secondly, an improved density peak clustering was performed based on the siphon balance rule. And beginning from the cluster centers, the unlabeled samples with high mutual neighbor degrees were selected in a progressive manner, then they were labeled by the two base classifiers. Finally, the final categories of the samples with inconsistent labels were determined by the shared subspace obtained by the multi-view non-negative matrix factorization algorithm. In the proposed algorithm, the unlabeled samples with better representation of spatial structure were selected by the improved density peak clustering and mutual neighbor degree, and the same sample labeled by different labels was revised via shared subspace, solving the low classification accuracy problem caused by sample misclassification. The algorithm was validated by comparisons in multiple experiments on 9 UCI datasets, and experimental results show that the proposed algorithm has the highest classification accuracy rate in 7 data sets, and the second highest classification accuracy rate in the other 2 data sets.

Key words: co-training; density peak clustering; siphon balance rule; shared subspace; mutual neighbor degree

0　引言

协同训练是基于不同视图训练两个分类器来互相标记样本以扩充训练集的一个过程,属于半监督学习[1-2]的范畴。由于协同训练一方面考虑到现实世界中数据集的分布情况,充分利用数据集中大量无标记样本和少量有标记样本;另一方面考虑到多视图数据的多态性、多源性、多描述性等特点,充分利用数据集中各视图间的相关信息,使得协同训练广泛应用于图像处理[3]、工业样品检测[4]、城市管理[5]、在线问答服务[6]等领域。

协同训练的特点是在不用人工干预的情况下,通过两个独立分类器的协同作用从大量无标记样本中自动获取训练信息;然而,在协同训练迭代过程中,多分类器存在的标记不一致样本易导致分类错误累积,而如何选择更具代表性的无标记样本加入训练集也是提高分类器性能的关键。针对无标记

收稿日期:2020-07-24;**修回日期:**2020-10-06;**录用日期:**2020-10-14。
基金项目:国家自然科学基金重大项目(11991024);重庆市高校创新研究群体项目(CXQT20015);重庆市研究生科研创新项目(CYS20241)。
作者简介:吕佳(1978—),女,四川眉山人,教授,博士,CCF会员,主要研究方向:机器学习、数据挖掘;　鲜焱(1995—),女,重庆涪陵人,硕士研究生,主要研究方向:机器学习、数据挖掘。

www.joca.cn

样本的选择,付治等[7]利用基于简单约束的聚类算法和标记置信度来选择可靠的实例,以此来实现有标记样本的扩展;然而该方法严重依赖标记置信度,若样本集的簇中心选取不准确,则会造成分类连带错误。Gan等[8]用半监督模糊聚类来发现数据的内部结构,利用该数据结构选取局部判别能力强的无标记样本,但该方法未考虑到数据的全局结构信息,使得选择的无标记样本不能充分代表数据的原始空间结构。Wu等[9]提出了一种自训练半监督分类框架,利用密度峰值聚类(Density Peak Clustering, DPC)算法[10]发现数据的空间结构后将其结构融入自训练迭代过程中。DPC算法适用于任意形状的数据集聚类,但需要通过决策图来选择簇中心,从而增加了算法的冗余性,限制了算法的应用领域。马春来等[11]针对DPC算法中对簇中心的判断问题,引入基于拐点法的簇中心选择策略来实现簇中心的自动选择,降低了通过决策图主观确定簇中心带来的误差。由于考虑到不同数据集稀疏或密集程度的差异,认为拐点位置无法为簇中心的定性划分提供理论依据,从而导致无法精确选取簇中心。龚彦鹭等[12]考虑到无标记样本隐含的空间结构信息,利用DPC算法选取局部密度与相对距离之和较高的无标记样本来辅助训练分类器,依据数据集全局的密集或稀疏程度来划分样本的选取区间,但忽略了样本的内部结构差异,无法体现样本的局部紧密或稀疏程度,可能会造成选取的无标记样本不可靠。

在分类器的协同过程中,多个分类器标记同一个无标记样本可能会产生类标号不一致的情况。为解决该问题,Zhao等[13]设计了基于伪标记误差的安全验证方法,在候选标记中,选择比前一轮迭代过程中误差精度低的标记作为无标记样本的最终类别,然而该方法通过伪标记误差来确定候选标记时,对误差较低的界限没有明确说明。龚彦鹭等[12]采用加权K近邻算法(Weight K-Nearest Neighbor, WKNN)为分类不一致的无标记样本重新进行分类,但该算法的分类正确率在数据集分布不均衡的情况下会受到影响。Peng等[14]提出多视图增强算法Boost. SH,通过利用多个视图的共享样本权重分布来保证视图的一致性,一定程度上解决了样本分类不一致问题,但由于多视图数据集的多态性等特点,该算法并不总是能找到满足条件的共享样本。

为了更好地选择隐含有用信息多的无标记样本和解决标记不一致样本的修订问题,提出了结合改进密度峰值聚类和共享子空间的协同训练算法(Co-Training algorithm combining Improved Density Peak Clustering and Shared Subspace, CTIDPCSS)。该算法通过定义虹吸平衡法则来实现簇中心的自动选择,以提高簇中心选择的准确度;通过引入相互邻近度[15],结合数据集全局和局部特征来划分样本的选取区间,以得到更具代表性的无标记样本;借助共享子空间(Shared Subspace, SS)[16-18],充分利用多视图间的一致性信息,得到一种新的样本修订方法。仿真实验结果表明,该算法具有更好的分类能力。

1　提出算法

在协同训练的迭代过程中,通常有两种方法选择无标记样本。一种是利用聚类算法来分析数据集的全局空间结构,根据样本所属类簇制定规则来划分样本选取区间,得到较好表现数据原始空间结构的样本,但该方法忽略了样本间的内部联系,易造成所选样本信息缺失。另一种是利用能体现样本间局部连带关系的算法,选取较好表现数据局部结构的样本,由于该方法忽略了数据的整体结构,易误将局部优选样本当作全局优选样本。若能够探讨一种将两者优势互补的方法,选择出更具代表性的无标记样本,则能提高分类器的分类性能。由此,本文考虑到DPC算法能探索到任何簇结构数据集的全局空间结构,而相互邻近度能体现样本的局部差异,故采用一种结合DPC和相互邻近度的方法来选择无标记样本。该方法在聚类阶段确定簇中心时,由于DPC算法提供的理论依据不足且未实现算法批量自动化应用,则本文通过定义虹吸平衡法则来实现簇中心的自动选择,为簇中心的定性划分提供理论依据。因此,将该方法中的DPC算法替换为基于虹吸平衡法则的密度峰值聚类算法,再根据簇中心计算样本的相互邻近度,将其作为无标记样本的选择依据。被挑选出的无标记样本能更好地体现数据的原始空间结构,则利用这些样本包含的有用信息来提高模型的分类性能。然而,在多分类器的协同过程中存在的标记不一致样本会造成分类错误累积,需及时实施样本修订策略才能有效避免标记错误样本的干扰。修订策略的目的是让多分类器最大化它们对无标记样本的一致性,从而借助无标记样本提升分类器的学习性能。因此,本文在保持原始数据的分类和决策能力的前提下找到数据一致性表征的共享子空间,以训练一个一致性分类器对多分类器标记不一致的样本重新分类。

1.1　改进的密度峰值聚类算法

1.1.1　DPC算法

DPC算法[10]是一种基于密度的聚类方法,首先计算每个样本的局部密度和相对距离,在密度和距离的决策图上通过人工选取簇中心,最后根据密度值将非簇中心点划分至所属类簇。对于给定的数据集$\boldsymbol{D}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbf{R}^{m\times n}$,其中$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{im}]$,$n$为样本总数,$m$为样本维度,计算每个样本$\boldsymbol{x}_i$的局部密度$\rho_i$和相对距离$\delta_i$。考虑到本文存在较大规模数据集,则局部密度$\rho_i$利用式(1)中的截断核[19]方式度量:

$$\rho_i=\sum_{i\neq j}\chi(d_{ij}-d_c);\ \chi(x)=\begin{cases}1, & x<0\\ 0, & x\geqslant 0\end{cases} \tag{1}$$

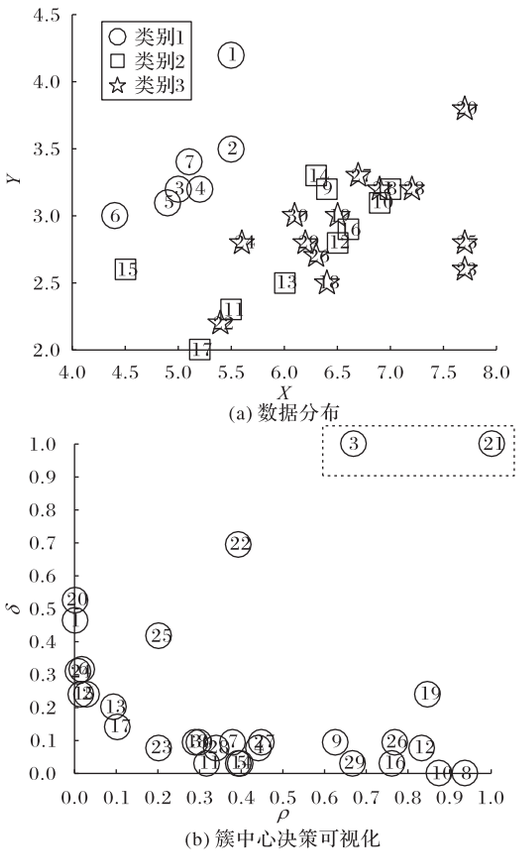
其中:局部密度ρ_i为以邻域截断距离d_c为半径的圆形结构内的样本数,d_{ij}为样本$\boldsymbol{x}_i$到$\boldsymbol{x}_j$的欧氏距离。

相对距离δ_i指比样本$\boldsymbol{x}_i$密度高且与它距离最近的样本$\boldsymbol{x}_j$的距离,计算公式如下:

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij}) \tag{2}$$

根据式(1)和式(2)的计算结果,以局部密度ρ_i为横轴,相对距离δ_i为纵轴作决策图,决策过程如图1所示。图1(a)是一个共含有30个样本的2维数据分布图,其类别分别以3种不同的图形来区分。图1(b)是DPC算法的可视化决策图,从图中"最上最右"的点出发,按照"向下向左"的原则,人工选择标号为3、21的点作为簇中心。但由于没有定量分析和决定性划分理论作为判断依据,人工选择簇中心存在一定主观性和误差。而作为簇中心的样本是影响DPC算法聚类效果的一个关键原因,若选取的簇中心不准确或错误,将会造成聚类效果不佳。

www.joca.cn



(a) 数据分布

(b) 簇中心决策可视化

图 1 DPC算法的簇中心决策图

Fig. 1 Cluster center decision diagram of DPC algorithm

1.1.2 基于虹吸平衡法则的DPC算法

针对DPC算法存在的不足，马春来等[11]从量化的角度引入簇中心权值的概念，提出根据权值的偏离度变化趋势定位拐点位置的簇中心选择策略，实现了簇中心的自动选择，但未提供簇中心决定性划分的理论依据。本文基于虹吸现象[20]提出虹吸平衡法则，为簇中心的选择提供定量分析和决定性理论依据。虹吸现象的原理是：当管状结构两端存在压力差（位能差）时，水将从压力大的一侧流向压力小的一侧，当两侧的压力相等时整个装置达到平衡状态，此时容器内的水面到达相同的高度，水就会停止流动。

在虹吸平衡法则中，为公开化平衡标准，将ρ_i和δ_i进行归一化后引入簇中心权值作为平衡量：

定义1 簇中心权值。

$$\gamma_i = \rho_i^* \delta_i \tag{3}$$

鉴于在一般情况下，数据集的簇类别数远小于样本总数，为了降低计算复杂度，将簇中心权值降序排列后取前q个点作为簇中心候选点，构造簇中心候选集$\boldsymbol{Q}_{1\times q}=\{\gamma_1,\gamma_2,\cdots,\gamma_q\}$。为引入位差，首先确定位差分割点$\gamma_c$，将点$\gamma_c$前的候选点分配到上位区$\boldsymbol{S}_{1\times c-1}=\{\gamma_1,\gamma_2,\cdots,\gamma_{c-1}\}$，点$\gamma_c$及其之后的候选点分配到下位区$\boldsymbol{X}_{1\times q-c+1}=\{\gamma_c,\gamma_{c+1},\cdots,\gamma_q\}$。

定义2 位差。

$$h(c)=\sum_{i=c}^{q}\gamma_i-\sum_{i=1}^{c-1}\gamma_i \tag{4}$$

其中：$\sum_{i=1}^{c-1}\gamma_i$表示上位区候选点的总簇中心权值，$\sum_{i=c}^{q}\gamma_i$表示下位区候选点的总簇中心权值，当存在自然的差距时，权值量以及与其相对应的簇中心候选点将会自发地从上位区转移到下位区。设置初始状态的位差分割点为$\gamma_c=\gamma_{q/2}$，每发生一次转移动作时位差分割点将向高权值位前进一步，直到到达相对平衡点则转移动作和前进动作均停止。

定义3 相对平衡点。

$$\arg\min_c h(c)=\{c|c\in[2,q/2],h(c)\geqslant 0\} \tag{5}$$

当位差分割点γ_c的下标取值范围为$[2,q/2]$，存在数值c使得$h(c)\geqslant 0$时，则认为簇中心权值达到平衡状态，此时停止簇中心权值和候选点的转移动作。从理论的角度分析，当少数簇中心候选点的簇中心权值总和与多数簇中心候选点的簇中心权值总和达到相对平衡时，认为少数高权值点能有效诠释多数低权值点，则基于权值来划分簇中心归属时多数与之差异度较大，从而可以将两者清楚地划分，以便选取到准确的簇中心。以下给出了基于虹吸平衡法则的密度峰值聚类算法的具体步骤，即改进的密度峰值聚类算法（Improved Density Peak Clustering，IDPC）。

输入：数据集$\boldsymbol{D}$，截断距离d_c，簇中心候选个数q；

输出：聚类结果。

过程：

步骤1 计算每个样本的局部密度ρ_i、相对距离δ_i和簇中心权值γ_i。

步骤2 将γ_i按降序排列，取前q个点作为簇中心候选点。

步骤3 根据公式$h(c)=\sum_{i=c}^{q}\gamma_i-\sum_{i=1}^{c-1}\gamma_i$计算位差。

步骤4 满足条件$\exists c\in[2,q/2],h(c)\geqslant 0$时达平衡状态，令平衡点$b=\arg\min_c h(c)$，取$b$之前的点$\{1,2,\cdots,b\}$作为簇中心。

步骤5 根据簇中心点、数据对象标号及密度边界阈值，将剩余点分配到各簇或边界域。

步骤6 直到所有样本都分配完毕，聚类过程结束。

1.2 相互邻近度

相互邻近度是由赵嘉等[15]定义的一种样本间邻近程度的度量准则，考虑到样本的内部结构差异，当两个样本的相互邻近度越高时，说明两个样本越接近，彼此的联系越紧密，从而更能反映出数据间真实的密集或稀疏关系。对于数据集$\boldsymbol{D}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbf{R}^{m\times n}$，$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{im}]$。当每个点的近邻数为$k$时，计算公式如下所示：

$$w_{ij}=\begin{cases}\mathrm{e}^{-\frac{d_{ij}^2}{\sigma^2}}, & j\in knn(i)\\ 0, & \text{其他}\end{cases} \tag{6}$$

$$deg_{i\to j}=\frac{1}{k+1}\sum_{v\in[knn(i),i]}w_{vj};\ i\neq j \tag{7}$$

$$A_{i,j}=deg_{i\to j}\cdot deg_{j\to i} \tag{8}$$

其中：$\sigma^2=\frac{1}{n\cdot k}\sum_{i=1}^{n}\sum_{j\in knn(i)}d_{ij}^2$，式(6)的$w_{ij}$为样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的邻近度，式(7)的$deg_{i\to j}$表示样本$\boldsymbol{x}_i$到$\boldsymbol{x}_j$的相对邻近度，式(8)的$A_{i,j}$表示样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的相互邻近度。

1.3 共享子空间

在协同训练迭代过程中，多分类器标记同一个样本会存在标记不一致的情况，如图2的样例所示。图2以两种颜色分别代表两种类别，分类器h_1和分类器h_2对样本C1、C3、C4、C5的分类标记均相同，而对样本C2的标记却不同。这种情况需及时实施样本修订策略，将多分类器对样本的标记一致性最大化，才能有效避免标记错误样本的干扰。基于此，充分利用

www.joca.cn

能表示多视图间一致性信息的共享子空间[16]来修订标记不一致样本。

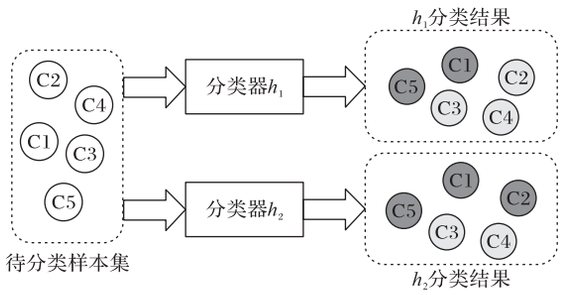


图2　多分类器标记不一致样例图

Fig. 2　Diagram of samples inconsistently labeled by multiple classifiers

共享子空间学习技术是指在保留原始数据的分类和决策能力的前提下,将原始多视图数据投影或因式分解得到一个共享低维子空间。非负矩阵分解(Non-negative Matrix Factorization, NMF)[21]是常用的因式分解方法,与其他矩阵分解方法的区别在于非负约束,有助于获得基于部分原始数据矩阵的表示,也增强了学习后的子空间的可解释性,但NMF只能处理单视图数据。Mekthanavanh等[22]利用多视图非负矩阵分解(Multi-view Non-negative Matrix Factorization, Multi-NMF)方法处理多视图数据,在保留原NMF算法优点的同时将其扩展到多视图领域。因此,本文通过在多视图有标记样本集中使用Multi-NMF算法因式分解得到一个相同的矩阵,该矩阵即为多视图有标记样本集的共享子空间。给定N个样本的非负多视图数据集$\{\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(n_v)}\}$,每个样本有$n_v$个视图。对于每个视图,$\boldsymbol{X}^{(v)}=\{x_1^{(v)},x_2^{(v)},\cdots,x_n^{(v)}\}\in\mathbf{R}^{D_v\times N}$,$\boldsymbol{U}^{(v)}$表示非负基矩阵,$\boldsymbol{V}^{(v)}$表示非负系数矩阵,使$\boldsymbol{X}\approx\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$。采用欧氏距离对Multi-NMF算法的目标函数进行求解,数学表达式为:

$$\min_{\boldsymbol{U}^{(v)},\boldsymbol{V}^{(v)},\boldsymbol{V}^*} L=\sum_{v=1}^{n_v}\left\|\boldsymbol{X}^{(v)}-\boldsymbol{U}^{(v)}(\boldsymbol{V}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\alpha_v\left\|\boldsymbol{V}^{(v)}-\boldsymbol{V}^*\right\|_{\mathrm{F}}^2$$

$$\text{s.t. }\boldsymbol{U}^{(v)},\boldsymbol{V}^{(v)},\boldsymbol{V}^*\geqslant 0,\forall 1\leqslant s\leqslant S,\left\|\boldsymbol{U}_{\cdot,s}^{(v)}\right\|=1 \tag{9}$$

其中:$\boldsymbol{V}^*$表示多个视图的共享矩阵,S为降维系数,α_v是控制不同视图重构误差项$\left\|\boldsymbol{X}^{(v)}-\boldsymbol{U}^{(v)}(\boldsymbol{V}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^2$与共享子空间约束项$\left\|\boldsymbol{V}^{(v)}-\boldsymbol{V}^*\right\|_{\mathrm{F}}^2$之间的相对权重。通过利用对角矩阵$\boldsymbol{Q}(v)=\left(\sum_{i=1}^{m_v}\boldsymbol{U}_{i,1}^{(v)},\sum_{i=1}^{m_v}\boldsymbol{U}_{i,2}^{(v)},\cdots,\sum_{i=1}^{m_v}\boldsymbol{U}_{i,S}^{(v)}\right)$简化$\boldsymbol{U}^{(v)}$上的约束条件,其中$m_v$为第$\boldsymbol{v}$个视图的维度。简化后的目标函数为:

$$\min_{\boldsymbol{U}^{(v)},\boldsymbol{V}^{(v)},\boldsymbol{V}^*} L=\sum_{v=1}^{n_v}\left\|\boldsymbol{X}^{(v)}-\boldsymbol{U}^{(v)}(\boldsymbol{V}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\alpha_v\left\|\boldsymbol{V}^{(v)}-\boldsymbol{V}^*\right\|_{\mathrm{F}}^2$$

$$\text{s.t. }\boldsymbol{U}^{(v)},\boldsymbol{V}^{(v)},\boldsymbol{V}^*\geqslant 0,\forall 1\leqslant \boldsymbol{v}\leqslant n_v \tag{10}$$

由于目标函数是非凸函数,不易求全局最优解,所以在求解过程中,采用逐次迭代更新方法[23]交替更新$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{V}^*$,直到目标函数收敛,最终得到共享子空间矩阵$\boldsymbol{V}^*$。

1.4　本文算法

基于以上分析,提出了结合改进密度峰值聚类和共享子空间的协同训练算法。所提算法模型的整体结构见图3,算法的具体过程如下所示。

输入:有标记样本集$\boldsymbol{L}$,无标记样本集$\boldsymbol{U}$,截断距离d_c,簇中心候选个数q,近邻数k,降维系数s,权重参数α_v;

输出:分类器h_1'、h_2'。

过程:

步骤1　在$\boldsymbol{L}$和$\boldsymbol{U}$上根据算法1进行IDPC聚类,找到簇中心$\{c_1,c_2,\cdots,c_b\}$。

步骤2　选择以簇中心$\{c_1,c_2,\cdots,c_b\}$为圆心,d_c为半径的圆形区域内的无标记样本,构造样本集$\boldsymbol{R}_1$。

步骤3　采用属性集合互补的方式将$\boldsymbol{L}$和$\boldsymbol{U}$划分为$\boldsymbol{v}_1$和$\boldsymbol{v}_2$,在$\boldsymbol{L}$上基于$\boldsymbol{v}_1$、$\boldsymbol{v}_2$分别训练出分类器h_1、h_2,用分类器h_1和h_2对样本集$\boldsymbol{R}_1$分类。

步骤4　在$\boldsymbol{L}$上利用Multi-NMF算法因式分解得到$\boldsymbol{v}_1$和$\boldsymbol{v}_2$的共享子空间$\boldsymbol{V}^*$,训练一致性分类器h^*。

步骤5　若h_1和h_2对样本的分类结果一致,则将该样本加入到$\boldsymbol{L}$集;若分类结果不一致,则用分类器h^*对其重新分类得到类标号后加入到$\boldsymbol{L}$集。

步骤6　$\boldsymbol{U}=\boldsymbol{U}-\boldsymbol{R}_1$,$\boldsymbol{L}=\boldsymbol{L}+\boldsymbol{R}_1$。

While $\boldsymbol{U}$集不为空

步骤7　计算$\boldsymbol{R}_1$集中每个样本与其k个邻居的相互邻近度,选择与$\boldsymbol{R}_1$集中每个样本相互邻近度最高的无标记样本,构造样本集$\boldsymbol{R}_2$。

步骤8　采用属性集合互补的方式将$\boldsymbol{L}$和$\boldsymbol{U}$划分为$\boldsymbol{v}_1'$和$\boldsymbol{v}_2'$,在$\boldsymbol{L}$上基于$\boldsymbol{v}_1'$、$\boldsymbol{v}_2'$分别得到分类器h_1'、h_2',用分类器h_1'和h_2'对样本集$\boldsymbol{R}_2$分类。

步骤9　在$\boldsymbol{L}$上利用Multi-NMF算法因式分解得到$\boldsymbol{v}_1'$和$\boldsymbol{v}_2'$的共享子空间$\boldsymbol{V}^{*\prime}$,训练一致性分类器$h^{*\prime}$。

步骤10　若h_1'和h_2'对样本的分类结果一致,则将该样本加入到$\boldsymbol{L}$集;若分类结果不一致,则用分类器$h^{*\prime}$对其重新分类得到类标号后加入到$\boldsymbol{L}$集。

步骤11　$\boldsymbol{U}=\boldsymbol{U}-\boldsymbol{R}_2$,$\boldsymbol{R}_1=\boldsymbol{R}_2$。

步骤12　$\boldsymbol{R}_2=\varnothing$。

End

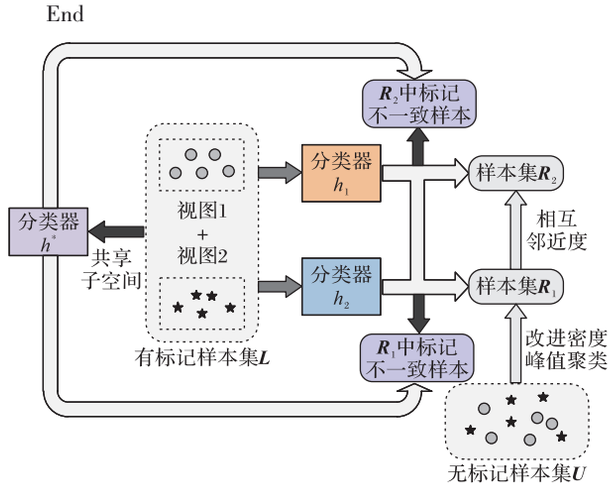


图3　本文算法结构

Fig. 3　Architecture of the presented algorithm

2　实验结果与分析

为了测试和评价本文算法的分类性能,将本文算法与4个对比算法在9个数据集上进行实验,计算出每种算法两个视图的平均分类正确率,作为评价本文分类模型的有效性指标。选用的对比算法如下:

1)标准协同训练算法(Co-Training, CT)。

2)结合文献[19]和共享子空间的协同训练算法(Co-Training algorithm combining Density Peak Clustering with Shared Subspace, CTDPCSS)。

www.joca.cn

3)文献[12]提出的算法 CTALDPCWKNN(Co-Training algorithm combined Active Learning and Density Peak Clustering and Weight *K*-Nearest Neighbor)。

4)结合文献[11]和共享子空间的协同训练算法(Co-Training algorithm combining Density Peak Clustering based on Cluster Center Automatic Selection Strategy with Shared Subspace, CTDPCCCASSSS)。

2.1 数据集描述

实验数据来源于UCI中的9个数据集。在实验过程中,5种算法在9个数据集上均采用十折交叉验证方法重复实验100次来测试算法的分类性能。每个数据集按照80%和20%的比例随机分为两部分,前者作为训练集用来构造分类模型,后者作为测试集用来测试模型的分类性能。其中,训练集依次随机选取10%、20%、30%、40%、50%的样本作为有标记样本,剩余样本去除标记作为无标记样本。表1描述了9个数据集的详细信息。

表1 数据集描述

Tab. 1 Description of datasets

Dataset	Instances	Features	Classes
Wine	178	13	3
Seeds	210	7	3
abalone	4 177	8	3
statlog	690	10	2
HTRU_2	17 898	8	2
Tic-Tac-Toe	958	9	2
ele	10 000	13	2
pima	768	8	2
Connectionist Bench	208	60	2

2.2 实验参数设置

本文依据文献[19]的经验法则将对比算法CTDPCSS、CTALDPCWKNN、CTDPCCCASSSS和本文算法需要设定的截断距离d_c赋值为2。CTALDPCWKNN算法还需设置模糊度参数ε,根据文献[12]设定ε=0.01。当上述算法利用Multi-NMF分解共享子空间矩阵时,设置权重参数α_v=0.01,降维系数s=0.5$d_{\min}$,其中$d_{\min}$表示多个视图中特征向量最少视图的特征维度。在不同规模的数据集中,相较于大样本数据集,数量上处于劣势的小样本数据集的分布比较稀疏,导致样本的局部差异度量尺度会有所不同,从而影响算法的参数设置。为了更客观地测试所设定的参数对本文算法的普适性,选择了3个具有代表性的数据集对簇中心候选个数q和近邻数k进行了参数调优,即分别在小、中、大样本量的数据集Wine、abalone和HTRU_2上进行性能测试,结果如图4所示。其中,近邻数k以1为步长,在[5,10]内选取;簇中心候选个数q以5为步长,在[10,30]内选取。

当有标记样本比例选取为10%时,图4说明了在3个不同规模的数据集上,参数值q、k对平均分类正确率的影响机制。由图4可知,当簇中心候选个数的取值范围为10≤q≤15或25≤q≤30,近邻数取值范围为5≤k≤8时,算法的平均分类正确率相对较高且处于稳定状态。当15<q<25时,由图4(a)和图4(b)发现,平均分类正确率有上升的波动,但随着k值的增加却逐渐降低,最终在一定范围内波动。在图4(a)中,当q=30,k=8时,存在一个明显的凸点,且该点的x轴和y轴坐标值在上述分类正确率较高的参数取值范围内。为了让本文算法的分类结果达到最优且保证算法的泛化性,在实验过程中将簇中心候选个数q和近邻数k分别统一设置为30和8。

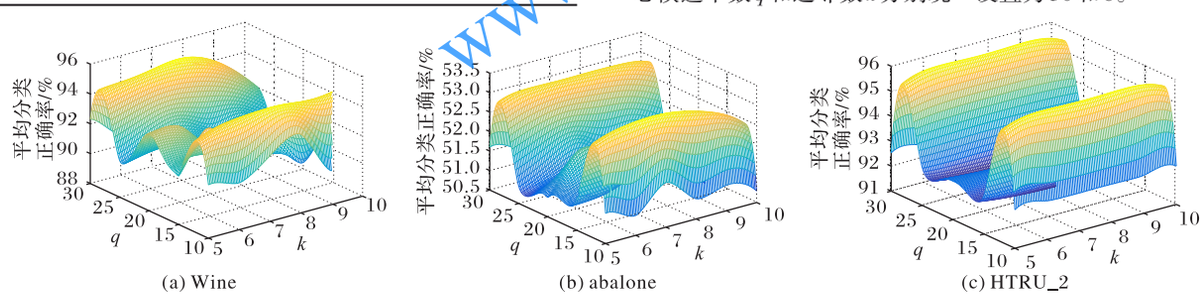


图4 参数q、k对平均分类正确率的影响

Fig. 4 Influence of parameters q and k on average classification accuracy

2.3 簇中心选取策略对比

为了验证本文提出的IDPC算法在选择簇中心时的准确性,以样本规模分别为小、中、大的数据集Wine、abalone和HTRU_2为例,分别采用基于决策图[19]、拐点法[11]和虹吸平衡法则的簇中心选择策略来选择簇中心。为了更直观地比较各簇中心选择策略的准确度,将每种方法得到的簇中心数与各数据集的类别数进行对比。在同一个数据集中,若各方法所得簇中心数越接近数据集本身的类别数,则表明该方法选取簇中心的准确度越高。在表2中,第5列代表各数据集本身的簇类别数,而第2、3、4列代表各簇中心选择策略分别在数据集Wine、abalone和HTRU_2下的簇中心数。

由表2可知,基于虹吸平衡法则的簇中心选择策略的准确度均高于基于决策图和拐点法的簇中心选择方法。原因在于利用决策图选择簇中心时主观性较强,当数据集的空间结构较密集时,簇中心与一般样本的相对决策位置差异不够明显,通过人眼无法精准地识别两者的黄金分割线,从而易造成簇中心数目多选或少选。在数据集稀疏程度不同的情况下,若簇中心候选集下位区中顺位靠后的任意两个相邻点,其簇中心权值的平均变化率大于上位区顺位靠前的两个相邻点,则利用拐点法来确定偏离度变化趋势的最大值点时,会造成簇中心数目多选。而虹吸平衡法则在不同稀疏程度或任意形状的数据集下,都能通过量化分析自动确定最优的簇中心数,从而有效降低了人工辅助选择簇中心带来的主观误差。

表2 不同选择策略的簇中心数与类别数对比

Tab. 2 Comparison of the number of cluster centers and the number of categories of different selection strategies

数据集	决策图	拐点法	虹吸平衡法则	类别数
Wine	2	3	3	3
abalone	5	4	3	3
HTRU_2	4	3	2	2

www.joca.cn

2.4　小样本数据集的处理

虹吸平衡法则是基于簇类别数远小于样本总数这一假设而提出,为了确定虹吸平衡法则在小样本数据集上的使用方法以及证明其在小样本数据集上的有效性,本文在Wine数据集上,分别选用其样本总数的20%、40%、60%、80%来构造新数据集。选取过程中,为了保证数据集的类别平衡,将每个类别的样本个数选取同等数量。在设定簇中心候选个数为30和将所有样本均纳入簇中心候选集这两种情况下,以平均分类正确率为评价指标,用构造的4个新数据集和原始Wine数据集作对比,实验结果如图5所示。

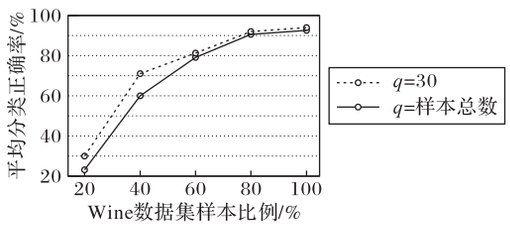


图5　小样本数据集上的平均分类正确率

Fig. 5　Average classification accuracy on small sample datasets

该实验过程中所得的簇中心数均为3,而由图5可知,随着数据量的增加,算法的平均分类正确率也逐渐上升,且q=30时的分类正确率均高于q=样本总数时的分类正确率,证明了虹吸平衡法则适用于各种规模的数据集寻找簇中心。因此,为了保持本文算法的分类稳定性,当数据集的样本总量小于30时,将其所有样本作为簇中心候选点,否则将簇中心候选个数统一设置为30。

2.5　算法分类正确率对比

为了说明本文算法的分类有效性,表3在有标记样本比例选取为10%时给出了5种算法在9个数据集上两个视图的平均分类正确率。

由表3可知,相较于4种对比算法,本文算法的平均分类正确率在9个数据集上最高提升了8.17个百分点。除了数据集Tic-Tac-Toe和pima外,本文算法在其余7个数据集上的分类性能均优于对比算法。因为基于IDPC算法选出的相互邻近度高的无标记样本更能代表数据集的原始空间结构,一定程度避免了边缘噪声点的影响,后续又通过共享子空间训练的一致性分类器修订样本,将多分类器对无标记样本的一致性最大化,最终得到较好分类效果的分类器。针对基于其他两种簇中心选择策略的密度峰值聚类算法,本文算法的正确率均高于基于决策图的CTDPCSS算法。因为决策图方法定位的簇中心不够准确,可能会造成聚类过程中部分样本的类簇分配错误,从而无法获得足够的有利于改善基分类器性能的无标记样本。除数据集Tic-Tac-Toe外,本文算法的正确率均高于基于拐点法的CTDPCCCASSSS算法,可能因为该数据集的权值偏离度变化趋势较明显,拐点位置很大概率会成为簇中心与一般样本的分界点。相较于采用加权KNN进行样本修订的CTALDPCWKNN算法,本文算法在数据集pima上的分类性能略低,这可能是因为该数据集本身具有模糊度高且易分的样本,故利用主动学习去标记模糊度高的样本适用此数据集。除pima数据集外,本文算法在其余8个数据集上的分类正确率均高于CTALDPCWKNN算法,证明了包含视图间一致性信息的共享子空间在重新分类标记不一致样本过程中的可行性,有效避免了由于标记不一致样本造成分类错误累积的问题。

www.joca.cn

表3　五种算法在9个数据集上的平均分类正确率对比　　单位:%

Tab. 3　Comparison of average classification accuracy of 5 algorithms on 9 datasets　　unit:%

Dataset	CT	CTDPCSS	CTALDPCWKNN	CTDPCCCASSSS	本文算法
Wine	92.021 2	91.522 4	90.746 0	92.807 2	**93.312 1**
Seeds	86.881 0	87.124 0	87.557 1	87.738 1	**87.881 0**
abalone	51.012 2	51.020 3	50.754 1	51.300 0	**51.500 4**
statlog	66.529 0	67.064 0	67.232 6	64.369 6	**68.144 9**
HTRU_2	91.878 8	92.110 1	92.351 4	92.621 8	**93.529 9**
Tic-Tac-Toe	57.072 9	63.318 0	64.262 0	**65.834 6**	65.243 6
ele	82.369 5	83.332 2	84.541 0	80.011 6	**87.921 4**
pima	65.645 3	66.656 0	**68.455 5**	67.448 4	67.824 5
Connectionist Bench	55.559 5	56.291 0	56.354 1	57.441 7	**57.902 4**

2.6　时间复杂度分析

为了分析5种算法的时间成本,本文通过划分每个算法的组成成分并计算各成分的时间复杂度,最终得到每个算法总的时间度复杂度,结果如表4所示。其中,各算法成分括号内的信息为该部分的时间复杂度。

由表4可知,本文算法和对比算法的时间复杂度均相同,则表明5种算法在解决模型分类问题时所需的大致时间相同。相较于4种对比算法,本文算法的组成成分略多,但由于各部分的时间复杂度均未超过$O(n^2)$,则总的时间复杂度保持不变。所以,本文算法在加入额外的优化步骤时,较好地控制了时间成本,保持了与对比算法同样的复杂度。

2.7　有标记样本的影响

在半监督学习中,有标记样本的比例会对模型的分类性能产生影响,为了讨论有标记样本的比例对算法的影响程度,图6给出了9个数据集在有标记样本比例分别为10%、20%、30%、40%、50%时的平均分类正确率曲线。

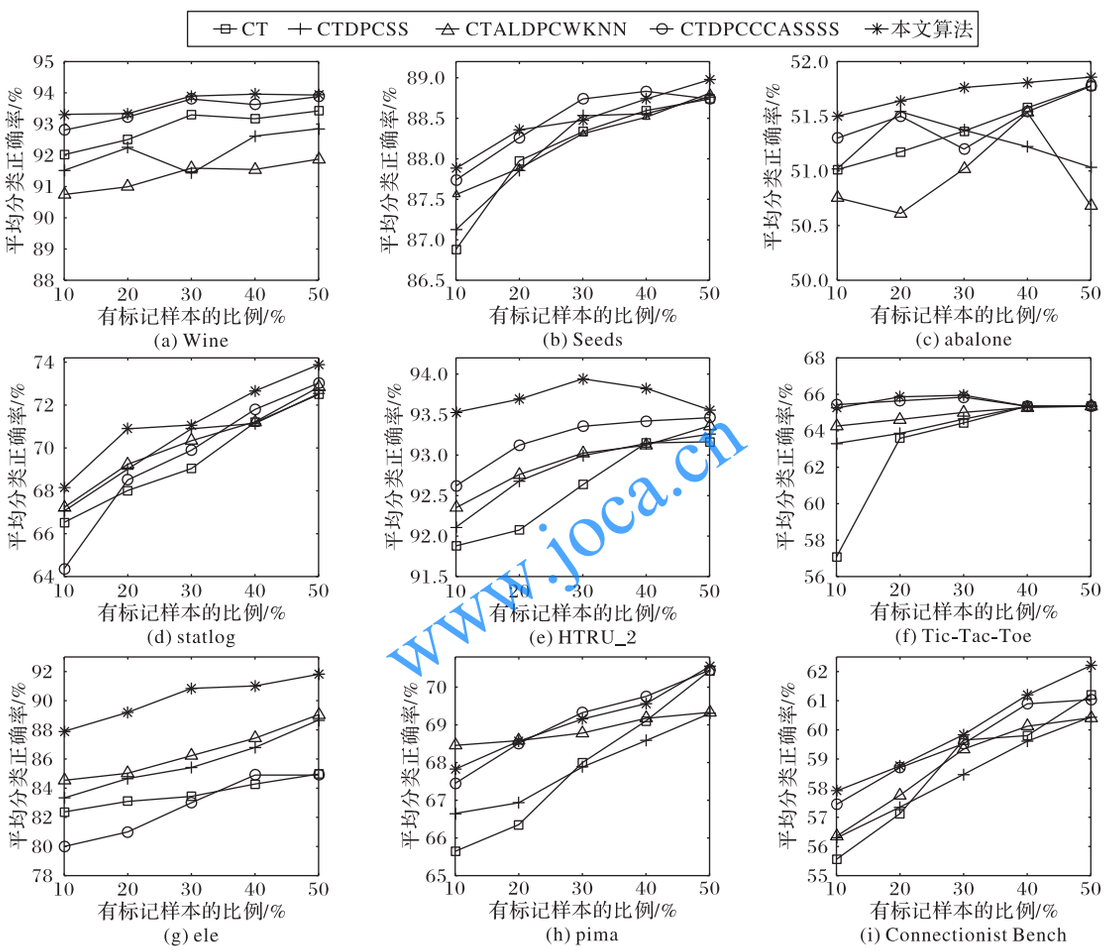
从图6总体的关系曲线分布来看,5种算法在除数据集HTRU_2和Tic-Tac-Toe外的其余7个数据集上,平均分类正确率随着有标记样本比例的增加而提升,这表明在半监督协同训练学习过程中,迭代初期有标记样本量越多则训练的基分类器越准确,进而利用分类器间的协同作用最终得到两个能独立完备确定样本类别的分类器。由图6(e)和图6(f)可知,在数据集HTRU_2和Tic-Tac-Toe中,有标记样本比例为10%~30%时的平均分类正确率呈上升趋势,然而当比例提高至40%~50%时,正确率下降且最终低于比例为10%时的正确率。因为随着有标记样本增加到一定比例后,有标记样本的

信息对分类器的影响逐渐减弱，从而达到一种相对稳定的状态，则此时的平均分类正确率会在一定范围内波动。

表 4 五种算法的时间复杂度对比

Tab. 4 Comparison of time complexity of 5 algorithms

算法名称	算法成分	总时间复杂度
CT	协同训练($O(n^2)$)	$O(n^2)$
CTDPCSS	协同训练($O(n^2)$)、密度峰值聚类($O(n^2)$)、共享子空间($O(n^2)$)	$O(n^2)$
CTALDPCWKNN	协同训练($O(n^2)$)、主动学习($O(n)$)、密度峰值聚类($O(n^2)$))、加权K近邻($O(n)$)	$O(n^2)$
CTDPCCCASSSS	协同训练($O(n^2)$)、拐点法簇中心选择策略($O(n)$)、密度峰值聚类($O(n^2)$)、共享子空间($O(n^2)$)	$O(n^2)$
本文算法	协同训练($O(n^2)$)、虹吸平衡法则簇中心选择策略($O(n)$)、密度峰值聚类($O(n^2)$)、相互邻近度($O(n^2)$)、共享子空间($O(n^2)$)	$O(n^2)$



www.joca.cn

图 6 有标记样本比例与平均分类正确率的关系曲线

Fig. 6 Relationship curve of labeled sample proportion and average classification accuracy

3 结语

本文针对协同训练迭代过程中用来改善分类器性能的无标记样本隐含有用信息不足和多分类器标记不一致样本的问题，提出了结合改进密度峰值聚类和共享子空间的协同训练算法。该算法通过定义虹吸平衡法则选择簇中心，再结合相互邻近度选择出综合数据全局和局部空间结构的无标记样本，最后利用共享子空间训练的一致性分类器重新分类标记不一致样本。基于虹吸平衡法则的改进密度峰值聚类算法为簇中心的选择提供定量分析和决定性理论依据，实现了簇中心的自动选择。通过将聚类算法与相互邻近度结合，综合考虑数据全局和局部的紧密和稀疏程度，选择更能代表数据原始空间结构的无标记样本。引入共享子空间，充分利用多视图间的一致性信息将标记不一致样本统一描述。在 9 个数据集上的仿真实验证明了该算法的有效性，实验结果表明该算法具有更高的模型评价能力。在后续的工作中，将进一步讨论算法的最优参数并给出具体的调优策略，以及探索本文算法的应用拓展性，如将本文算法应用于多标签分类问题。

参考文献（References）

[1] 屠恩美，杨杰. 半监督学习理论及其研究进展概述[J]. 上海交通大学学报，2018，52(10)：1280-1291.（TU E M，YANG J. A review of semi-supervised learning theories and recent advances [J]. Journal of Shanghai Jiao Tong University，2018，52(10)：1280-1291.）

[2] GOUTTE C，CANCEDDA N，DYMETMAN M，et al. Semi-supervised learning for machine translation[J]. Journal of the Royal

Statistical Society, 2017, 172(2): 530-530.

[3] HU T, HUANG X, LI J, et al. A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery [J]. Remote Sensing of Environment, 2018, 217: 144-157.

[4] LI X, LU H, YANG J, et al. Semi-supervised LIBS quantitative analysis method based on co-training regression model with selection of effective unlabeled samples [J]. Plasma Science and Technology, 2019, 21(3): No. 034015.

[5] XU Z, CAO Y, KANG Y. Deep spatiotemporal residual early-late fusion network for city region vehicle emission pollution prediction [J]. Neurocomputing, 2019, 355: 183-199.

[6] ZHANG Z, HU Z, YANG H, et al. Factorization machines and deep views-based co-training for improving answer quality prediction in online health expert question-answering services [J]. Journal of Biomedical Informatics, 2018, 87: 21-36.

[7] 付治,王红军,李天瑞,等. 基于k个标记样本的弱监督学习框架[J]. 软件学报, 2020, 31(4): 981-990. (FU Z, WANG H J, LI T R, et al. Weakly supervised learning framework based on k labeled samples[J]. Journal of Software, 2020, 31(4): 981-990.)

[8] GAN H, LUO Z, FAN Y, et al. Enhanced manifold regularization for semi-supervised classification[J]. Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2016, 33(6): 1207-1213.

[9] WU D, SHANG M, LUO X, et al. Self-training semi-supervised classification based on density peaks of data[J]. Neurocomputing, 2018, 275: 180-191.

[10] 陈叶旺,申莲莲,钟才明,等. 密度峰值聚类算法综述[J]. 计算机研究与发展, 2020, 57(2): 378-394. (CHEN Y W, SHEN L L, ZHONG C M, et al. Survey on density peak clustering algorithm [J]. Journal of Computer Research and Development, 2020, 57(2): 378-394.)

[11] 马春来,单洪,马涛. 一种基于簇中心点自动选择策略的密度峰值聚类算法[J]. 计算机科学, 2016, 43(7): 255-258, 280. (MA C L, SHAN H, MA T. Improved density peaks based clustering algorithm with strategy choosing cluster center automatically [J]. Computer Science, 2016, 43(7): 255-258, 280.)

[12] 龚彦鹭,吕佳. 结合主动学习和密度峰值聚类的协同训练算法[J]. 计算机应用, 2019, 39(8): 2297-2301. (GONG Y L, LYU J. Co-training algorithm with combination of active learning and density peak clustering [J]. Journal of Computer Applications, 2019, 39(8): 2297-2301.)

[13] ZHAO J, LIU N, MALOV A. Safe semi-supervised classification algorithm combined with active learning sampling strategy [J]. Journal of Intelligent and Fuzzy Systems, 2018, 35(4): 4001-4010.

[14] PENG J, AVED A J, SEETHARAMAN G, et al. Multiview boosting with information propagation for classification [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(3): 657-669.

[15] 赵嘉,姚占峰,吕莉,等. 基于相互邻近度的密度峰值聚类算法[J/OL]. 控制与决策, [2020-06-13]. https://doi.org/10.13195/j.kzyjc.2019.0795. (ZHAO J, YAO Z F, LYU L, et al. Density peaks clustering based on mutual neighbor degree[J]. Control and Decision, [2020-06-13]. https://doi.org/10.13195/j.kzyjc.2019.0795.)

[16] TAN Q, YU G, WANG J, et al. Individuality- and commonality-based multiview multilabel learning [J]. IEEE Transactions on Cybernetics, 2019(Early Access): 1-12.

[17] YUAN H, TANG Y Y. Spectral-spatial shared linear regression for hyperspectral image classification [J]. IEEE Transaction on Cybernetics, 2017, 47(4): 934-945.

[18] XU N, GUO Y, WANG J, et al. Multi-view clustering via simultaneously learning shared subspace and affinity matrix [J]. International Journal of Advanced Robotic Systems, 2017, 14(6): 1-8.

[19] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[20] JI D Y, KIM S H, LEE K Y, et al. Experimental study of small scale siphon breaker to verify Siphon Breaker Simulation Program (SBSP) [J]. Annals of Nuclear Energy, 2018, 121: 406-413.

[21] JIANG W, MA T, FENG X, et al. Robust semi-nonnegative matrix factorization with adaptive graph regularization for gene representation[J]. Chinese Journal of Electronics, 2020, 29(1): 122-131.

[22] MEKTHANAVANH V, LI T, MENG H, et al. Social Web video clustering based on multi-view clustering via nonnegative matrix factorization [J]. International Journal of Machine Learning and Cybernetics, 2019, 10(10): 2779-2790.

[23] CAI H, LIU B, XIAO Y, et al. Semi-supervised multi-view clustering based on orthonormality-constrained nonnegative matrix factorization[J]. Information Sciences, 2020, 536: 171-184.

This work is partially supported by the Major Project of the National Natural Science Foundation of China (11991024), the Program of Chongqing University Innovation Research Group (CXQT20015), the Chongqing Graduate Research and Innovation Project (CYS20241).

LYU Jia, born in 1978, Ph. D., professor. Her research interests include machine learning, data mining.

XIAN Yan, born in 1995, M. S. candidate. Her research interests include machine learning, data mining.

www.joca.cn