DOI:10.11721/cqnuj20210404

基于加权主成分分析和改进密度峰值聚类的协同训练算法*

龚 旭,吕 佳
(重庆师范大学 计算机与信息科学学院 重庆市数字农业服务工程技术研究中心,重庆 401331)

摘要:【目的】针对协同训练算法在视图分割时未考虑噪声影响和两视图分类器对无标记样本标注不一致问题,提出了基于加权主成分分析和改进密度峰值聚类的协同训练算法。【方法】首先引入加权主成分分析对数据进行预处理,通过寻求初始有标记样本中特征和类标记之间的依赖关系求得各特征加权系数,再对加权变换后的数据进行降维并提取高贡献度特征进行视图分割,这一策略能较好地过滤视图分割时引入的噪声,同时保证数据中的关键特征能均衡划分到两个视图,从而更好地实现两个分类器的协同作用;同时,在密度峰值聚类上提出一种"双拐点"法来自动选择聚类中心,利用改进后的密度峰值聚类来确定标记不一致样本的最终类别,以降低样本被误分类的概率。【结果】与对比算法相比,所提算法在分类准确率和算法稳定性上有较大提升。【结论】通过加权主成分分析能有效地过滤掉视图分割中的噪声特征,同时改进后的密度峰值聚类减少了样本被误标记的概率。

关键词:协同训练;加权主成分分析;密度峰值聚类;"双拐点"法
中图分类号:TP18 **文献标志码**:A **文章编号**:1672-6693(2021)04-0087-10

多视图数据[1]广泛存在于现实世界中,它从不同角度对目标对象进行描述和表征,通过分析和利用不同视图的互补信息对目标更详尽地学习。因此,为了充分利用多视图数据中蕴含的信息,多视图学习(Multi-view learning,MVL)[2]应运而生。多视图学习算法分为子空间学习、多核学习和协同训练[3-4]这 3 类,其中协同训练因为较好的性能和鲁棒性被研究者们关注,目前被广泛应用在图像处理、神经机器翻译、视频标签、自然语言处理和数字图像处理等领域。

协同训练的假设条件是数据存在两个充分冗余视图,但现实世界中大部分数据不存在这样两个既定的视图,如何有效地进行视图分割[5]成为提升协同训练算法性能的关键。Tang 等人[6]随机将样本特征划分为主视图和特权视图,提出一种全新的特权 SVM 分类模型 PSVM-2V,充分利用两种视图下的互补信息构建一个高效的分类器。文献[7]提出了一种不完整的多视图通用聚类框架,利用低秩表示法来发现数据的固有子空间结构进而构造每个视图,但基于谱约束实现每个视图的低维表示可能会引入特征噪声[8]。特征噪声是在分类识别过程中引入误差的特征,而协同训练算法在视图分割时常常忽略这些特征带来的影响进而导致训练过程中分类器性能降低。文献[8]表明就识别精度而言基于集成技术的噪声处理方法优于基于单一技术的方法,而后者在效率方面表现更好。Liang 等人[9]通过引入噪声鲁棒模块提出了一种新的噪声特征提取方法,实验表明基于该方法对噪声特征进行处理能有效提高分类精度。Zhou 等人[10]提出来一种基于特征冗余最小化的多视图潜空间学习框架,以希尔伯特-施密特独立性准则为约束条件在潜空间对样本数据进行重建后,在一定程度降低视图中噪声的影响,通过实验证明了该方法的有效性。Zhao 等人[11]则提出一种新颖的神经网络用于人脸识别,基于主成分分析(Principal component analysis, PCA)来减少图像数据中深层特征维数,消除其中冗余和被污染的视觉特征,结果表明在识别精度和效率上有极大的提升。但传统 PCA 在降维和去噪过程中对数据的所有特征同等对待,这一做法与实际情况不符,因为在分类识别中不同特征的贡献度和重要程度不同。加权主成分分析(Weight principal component analysis,WPCA)则是在 PCA 的基础上引入特征加权[12-13],为每一个特征设置加权系数来表示权重,在降维过程中选择高权重的特征作为关键特征。文献[14]通过最小二乘法线性拟合特征和类标记之

* **收稿日期**:2020-11-24 **修回日期**:2020-12-02 **网络出版时间**:2021-06-30 09:36
资助项目:国家自然科学基金(No. 11971084);重庆市教育委员会科技创新项目(No. KJCX220024);重庆市高校创新研究群体(No. CXQT20015);重庆市研究生科研创新项目(No. CYS20241)
第一作者简介:龚旭,男,研究方向为机器学习、数据挖掘,E-mail:gongxu@cqnu.edu.cn;**通信作者**:吕佳,女,教授,博士,E-mail:lvjia@cqnu.edu.cn
网络出版地址:https://kns.cnki.net/kcms/detail/50.1165.N.20210629.1802.008.html

间的依赖关系求得加权系数,在加权变换后的数据上进行降维。Tian 等人[15]则利用 WPCA 优化化工过程中的报警阈值,有效地降低了虚警率。文献[16]在线性判别分析(Linear discriminant analysis,LDA)中引入 WPCA 进行数据预处理,降低数据维度同时剔除引入噪声特征,实验表明该方法的分类性能有较大提升。

密度峰值聚类算法[17-18](Density peak clustering,DPC)是一种基于样本密度和距离发现簇的新型聚类算法,具有无须迭代求解、鲁棒性强和调节参数少等优点。但聚类中心的选择需要人工参与,在不同数据集上的自适应性差,算法性能不稳定。为此,马春来等人[19]则将密度和距离的乘积作为簇中心权值,根据权值的变化趋势搜索"拐点"以此来自动选择聚类中心。但该方法易将离群点误选为聚类中心,导致聚类过程中引入误差。文献[20]则利用密度敏感的相似性在流形数据集中寻找聚类中心,减少参数对聚类结果的影响,同时设计了一种新的密度聚类指数 DCI 代替决策图来自动确定聚类中心的数目,改进后的算法在合成的文本数据集和人脸数据库取得较好的识别效果。在文献[17]中密度峰值聚类同样被用于人脸数据库识别,但人为选择聚类中心导致算法性能波动较大。Wang 等人[21]使用改进后的密度峰值聚类算法对 IBP 数据进行 5 种血压类型识别,实验表明该算法性能可与最新的无监督图像分类技术媲美。文献[18-21]表明基于改进聚类中心选择后的密度峰值聚类在无标记样本上聚类得到的簇对于簇中样本的真实类别具有指示作用。当两个分类器对同一样本标记产生分歧时,引入密度峰值聚类确定该样本所属簇,基于同一簇中样本标记的大多数对标记不一致样本进行修正,能降低样本被误标记概率进而提高分类器性能。

由于加权主成分分析能有效过滤数据中的特征噪声并能基于特征贡献度提取出关键特征,因此本文将之引入传统协同训练算法来解决视图分割不均和特征噪声问题;而改进后的密度峰值聚类对于可能被误分类样本有较好的修正作用,能减少迭代过程中错误的累积,提升分类器性能。因此,本文将提出基于加权主成分分析和改进密度峰值聚类的协同训练算法,并通过 4 组实验对算法的分类精度和执行速度进行验证。

1 提出算法

在协同训练算法中进行视图分割时常会忽略数据中存在的噪声和冗余特征,噪声会在迭代训练中造成误差的累积加大,而冗余特征并不包含分类所需的关键信息,反而会增加数据维度降低算法的效率。因此,考虑噪声和冗余特征将会有效地改善算法性能。WPCA 针对传统主成分分析的不足引入特征权值系数来表示各特征的重要程度,特征权值越大特征越重要、蕴含分类所需关键信息越多;相反,权值越小可认为该特征是产生干扰信息的噪声特征,在数据降维过程中可以剔除。

本文利用 WPCA 对数据进行预处理,根据降维后特征的贡献度进行视图分割,既达到剔除特征噪声和降维的目的,同时也能保证关键特征均衡划分到两个视图。此外,鉴于密度峰值聚类在分类识别中的良好表现,将分类不一致的样本放入到已标记样本集中聚类,基于同一簇中样本相似度高则同属一个类别概率较大的原则,用簇中样本类标记的大多数进行修正,降低样本被误标记的概率,从而有效提高分类器性能。

1.1 基于 WPCA 算法的降噪和视图分割

1.1.1 WPCA 算法 加权主成分分析是在传统 PCA 算法的基础上引入特征加权[12-13],为每一维特征增加一个加权系数来表示它的重要程度,值越大越重要,值越小则容易引入噪声,计算公式为:

$$\begin{cases} y_1 = w_1 x_{11} + w_2 x_{12} + \cdots + w_n x_{1n} \\ y_2 = w_2 x_{21} + w_2 x_{22} + \cdots + w_n x_{2n} \\ \qquad \cdots \\ y_m = w_1 x_{m1} + w_2 x_{m2} + \cdots + w_n x_{mn} \end{cases}。 \tag{1}$$

其中:$w_j(j=1,2,\cdots,n)$表示加权系数,$y_i(i=1,2,\cdots,m)$为样本类别,n 为特征数,m 为样本个数。使用最小二乘法来拟合样本和标记的依赖关系,w_j 的最优解通过

$$\min \sum_{i=1}^{m} (y_i - x_i \boldsymbol{w}^{\mathrm{T}})^2 \tag{2}$$

求得。随后,基于加权系数 w_j 加权变换后的数据由下式求得:

$$\boldsymbol{Z}_{m\times n} = \begin{bmatrix} w_1 x_{11} & \cdots & w_n x_{1n} \\ w_2 x_{21} & \cdots & w_n x_{2n} \\ \vdots & \cdots & \vdots \\ w_m x_{m1} & \cdots & w_n x_{mn} \end{bmatrix}。 \tag{3}$$

至此，得到加权变换后的数据，在此基础上再对 $\boldsymbol{Z}_{m\times n}$ 进行降维。

首先对 $\boldsymbol{Z}_{m\times n}$ 进行零均值化处理，求得协方差矩阵：

$$\boldsymbol{c}=\frac{1}{m-1}\sum_{i=1}^{m}(z_i-\boldsymbol{a})^{\mathrm{T}}(z_i-\boldsymbol{a}), \tag{4}$$

其中：z_i 表示第 i 个样本，$\boldsymbol{a}$ 是对 $\boldsymbol{Z}_{m\times n}$ 零均值化的矩阵。根据

$$\boldsymbol{cu}=\lambda\boldsymbol{u} \tag{5}$$

求出 $\boldsymbol{c}$ 的特征值 $\lambda_i(i=1,2,\cdots,n)$ 和特征向量 $\boldsymbol{u}_i(i=1,2,\cdots,n)$。

将降序排列的特征值对应的特征向量组合成为原始数据的变换矩阵 $\boldsymbol{p}=[u_1\quad u_2\quad\cdots\quad u_n]$，进而得到变换后的数据矩阵：

$$\boldsymbol{Y}=(\boldsymbol{Z}-\boldsymbol{a})\boldsymbol{p}。 \tag{6}$$

此时还需从 $\boldsymbol{Y}$ 中选择相应维数特征，为此每一分量定义一个贡献度，根据贡献度的高低选择特征，其中：y_k 表示 $\boldsymbol{Y}$ 中第 k 维特征贡献度，即：$y_k=\dfrac{\lambda_k}{\sum_{j=1}^{n}\lambda_j}$。

1.1.2 视图的分割　基于前文得到每个特征的贡献度 y_k，为保证降维后原始数据中关键信息得以保留，同时又能有效过滤特征噪声，设定贡献度阈值 t，当多个特征贡献度之和不小于 t 时，此时得到 k 值就是保留的特征数：

$$\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{n}\lambda_j}\geqslant t。 \tag{7}$$

根据降维后的数据，将 1 到 k 中奇数位特征划分到视图 v_1，偶数位特征划分到视图 v_2，完成视图的分割。

1.2 改进的密度峰值聚类修正样本标记

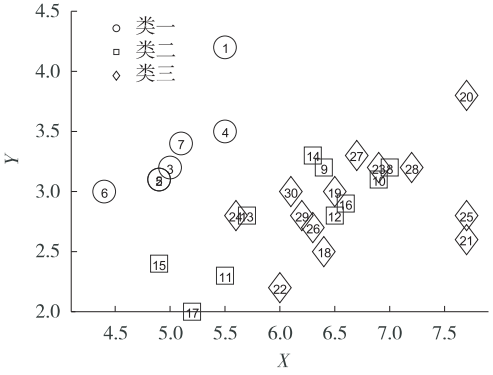
1.2.1 密度峰值聚类修正样本标记原理　密度峰值聚类首次在文献[17]中提出，是一种新型聚类算法，核心思想在于对聚类中心的刻画。聚类中心应同时满足两个条件：1）自身密度要比周围样本点的密度大；2）与密度更大的样本点之间相对距离较远。设待聚类数据集为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i\in\mathbf{R}_d$，$i=1,2,\cdots,n$。样本点 x_i 的局部密度 ρ_i 采用 Cut-off Kernel 方法计算：$\rho_i=\sum_{i\neq j}\chi(d_{ij}-d_c)$，$\chi(x)=\begin{cases}1,x<0\\0,x\geqslant 0\end{cases}$。其中：$d_{ij}$ 表示样本点 x_i 和 x_j 的距离，本文采用欧氏距离来计算；d_c 作为截断距离需要人为指定。从 Cut-off Kernel 计算方法可以看出 ρ_i 是离散值，容易导致不同的样本点产生相同的密度，为了避免该问题发生，使用 Gaussian Kernel 代替 Cut-off Kernel 方法计算局部密度：$\rho_i=\sum_{i\neq j}\mathrm{e}^{-\left(\frac{d_{ij}}{d_c}\right)^2}$，样本点 x_i 的相对距离为 $\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}(d_{ij})\\\max\limits_{i:\rho_i=\max(\rho)}(\delta)\end{cases}$。

求得各点的局部密度和相对距离后，以 ρ_i 为横坐标，δ_i 为纵坐标画出决策图，然后人工选择相对距离大且局部密度高的样本点作为聚类中心，再将其余点归属到各个中心完成聚类。为了更好地理解密度峰值聚类对标记不一致样本的修正过程，以一个例子加以说明。图 1a 是一组 2 维数据的分布图，共包含 30 个样本点，其中存在 3 个类别，用 3 种不同图形来表示不同类别的样本点。图 1b 是密度峰值聚类的决策图，依据决策图人工选择标号 2，23，19，26 的点作为聚类中心，最后得到聚类结果如图 1c 所示。

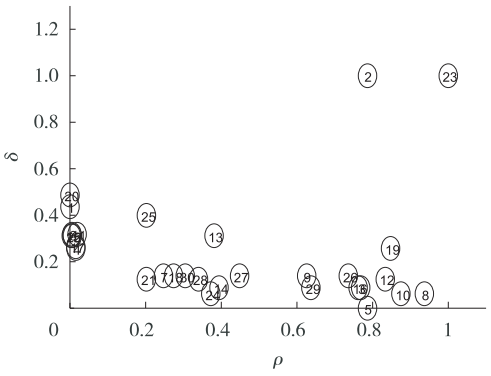
假设在分类过程中分类器对标号 20 的点标记不一致，此时引入密度峰值聚类找到该点所在簇，基于同一簇中相似度大，同属于一个类概率高的原则，将簇中样本类别的大多数作为修正后标记。从图 1c 中可以看出，该点所属簇共有 8 个样本，其中 6 个为类三，余下 2 个属于类二，故标号为 20 的样本最终会被正确修正为类三。实际上该簇中属于类三的样本都会被正确修正。当然错误修正情况会出现在样本点 8 和 10 上，但前提是分类器标记产生分歧。总体来看正确修正的概率远大于错误修正。在后面的实验 4 中也证明，在大多数数据集上对样本的修正是合理可行的。但修正的准确率受聚类效果影响，所以聚类中心的选择对于这一修正过程尤为关键。而上述算法中聚类中心的选择需要人工参与，存在一定主观性和误差，下文中将提出一种新的聚类中心选择方法。

1.2.2 “双拐点”聚类中心选择法　文献[17]中给出了自动选择聚类中心的思路，根据 $\rho_i\times\delta_i$ 的值来进行聚类中心选择，而大部分改进的方法也是基于此思路。文献[19]中提出一种拐点法来自动选择聚类中心，将 $\rho_i\times\delta_i$ 作为聚类中心的权值，选取前 30 个样本点，以样本点标号为横坐标，降序排列后的权值作为纵坐标画出决策图。

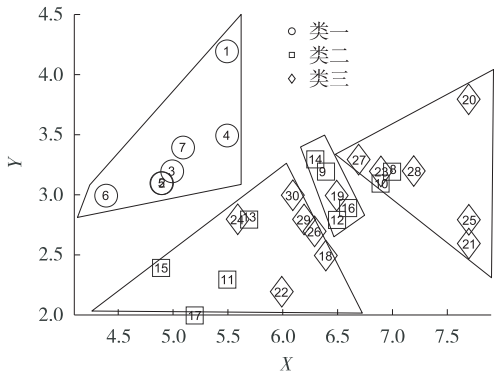
由于非聚类中心过渡到聚类中心时权值会出现跳跃,所以根据决策图中权值的变化趋势找出一个拐点(拐点即斜率最大处),将拐点之前的样本点作为聚类中心。然而,被选为聚类中心的样本点 x_i 的 $\rho_i \times \delta_i$ 的值可能包含 ρ_i 值较小而 δ_i 较大或 δ_i 值较小而 ρ_i 较大这两种情形,而实际上这类点可能是离群点。



a　样本点分布图



b　决策图



c　聚类结果图

图1　密度峰值聚类修正样本实例

Fig. 1　Example of density peak clustering correction sample

基于以上问题,受文献[19,21-22]的启发,在沿用自动选择聚类中心的思路上,本文提出一种"双拐点"聚类中心选择法。"双拐点"法认为同非聚类中心相比,聚类中心在 ρ_i 和 δ_i 上同时存在值的跳跃,因此不再将 $\rho_i \times \delta_i$ 的值作为中心权值来寻找拐点,而是在 ρ_i 和 δ_i 中分别寻找两个拐点 ρ_s 和 δ_s,然后基于 ρ_s 和 δ_s 的值来确定聚类中心的边界,选取密度大于 ρ_s 且距离大于 δ_s 点作为聚类中心。在选择聚类中心过程中,"双拐点"法综合考虑到距离和密度的变化,有效避免离群点带来的影响,同时达到自动选择聚类中心的目的。图2给出了基于"双拐点"法的密度峰值聚类结果图,其中标号为2,23,19,26和12的点被选为聚类中心。与图1c中结果相比,以标号2,26为中心的簇相同。

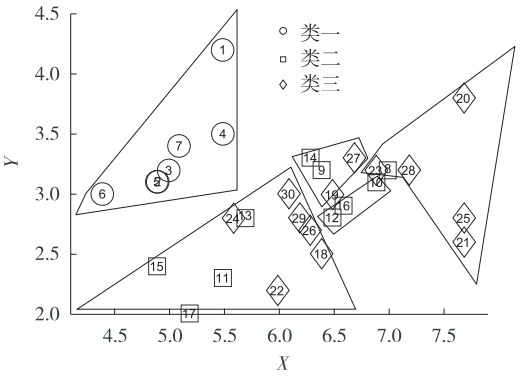


图2　"双拐点"法聚类结果图

Fig. 2　Clustering result of "double turning point" method

在图2中增加了一个以标号12为中心的簇且簇中3个样本同属于类二,而标号20的点所在簇中仅存在一个样本点不属于类三。所以当多个分类器对标号为20的点标记不一致时,基于簇中样本标记的大多数能将该样本正确修正为类三。从图1c和图2来看,基于"双拐点"法的聚类效果要优于传统密度峰值聚类,且"双拐点"法能适应不同数据集的变化来自动选择聚类中心,对样本的修正取得了较好的效果。

1.3 本文算法流程

基于以上工作,本文提出了基于加权主成分分析和改进密度峰值聚类的协同训练算法,具体流程如下。

输入：有标记样本集 L，无标记样本集 U，每次迭代选取的无标记样本数 u，特征贡献度阈值 t，降维后数据维数 k，截断距离 d_c。

输出：分类器 h_1，h_2。

过程：

在 L 上利用(1)～(3)式计算出数据特征的加权系数，并得到加权变换后的数据 Z；

利用(4)～(6)式对加权变换后的数据进行降维，得到变换后低维空间上的数据；

根据贡献度阈值 t，利用(7)式求得 k，保留降维后数据的前 k 维特征，将奇数位特征划分到视图 v_1，偶数位特征划分到视图 v_2；

训练初始分类器：$(L, v_1) \rightarrow h_1$，$(L, v_2) \rightarrow h_2$；

while U 集不为空

　从 U 中随机选取 u 个样本构成子集 U'，不足 u 个则选择 U 中全部样本；

　while U' 集中样本未标记完

　　用分类器 h_1，h_2 对 U' 中样本进行标记；

　　if h_1，h_2 对样本标记不一致

　　使用改进的密度峰值聚类修正该样本标记；

　　end if

　　更新 U'；

　end while

　更新样本集：$L=L+U'$，$U=U-U'$，更新分类器 h_1，h_2；

end while

2 实验仿真

为了验证本文算法的性能，选用以下 4 个算法作为对比算法：

1) 标准协同训练算法(Co-training，CT)。

2) 基于文献[3]提出的结合半监督聚类和加权 KNN 的协同训练方法(Co-training method combined semi-supervised clustering and weighted K-nearest neighbor，CTSMCKNN)。

3) 在标准协同训练的基础上引入 DPC 算法提出的结合密度峰值聚类的协同训练方法(Co-training method combined with density peak clustering，CTDPC)。

4) 结合主动学习和密度峰值聚类的协同训练算法[23](Co-training algorithm combined with active learning and density peak clustering，CTALDPC)。

2.1 数据集和实验参数设置

为了保证实验的可靠性，实验中使用十折交叉验证来测试算法的分类准确率，采用 50 次实验的平均值作为最终结果，所有算法均使用朴素贝叶斯作为两个视图的基分类器。从训练集中随机选取 90%的样本擦除标记作为无标记样本集 U，其余样本保留标记作为有标记样本集 L。相关实验参数具体设置如下，加权 KNN 中最邻近参数设置为 5，密度峰值聚类的截断距离 $d_c=2$，每次迭代选择的无标记样本数 $u=30$，特征贡献度阈值 $t=0.95$。

实验数据集采用 UCI 数据库中的 9 个数据集，详细描述见表 1。

表 1　UCI 数据集描述

Tab. 1　UCI dataset description

数据集	大小	贡献度	分类	数据集	大小	贡献度	分类
Banknote Authentication	1 372	4	2	Eegeyestaage	14 980	15	2
Abalone	4 177	8	3	Breast Cancer	699	10	2
Contraceptive	1 473	9	3	Spectf Heart	267	44	2
Wine	178	13	3	Connectionist Bench	208	60	2
Liver Disorders	345	6	2				

2.2 实验1

本次实验是为了研究当初始有标记样本数量较少时,算法的性能表现。表2给出了初始有标记样本比例为10%时,两视图分类器h_1和h_2的分类正确率,表3则给出了最终分类器的分类正确率。

由表2可知,在数据集Connectionist BenchBanknote、Contraceptive和Spectf Heart上,本文算法的两个视图分类器性能均高于对比算法。在Wine、Liver Disorders和Breast Cancer这3个数据集中,存在一个视图分类器性能强于对比算法,而另外一个分类器表现差异不大。这表明基于WPCA处理噪声的有效性,经加权系数变换后的数据充分考虑到每一维特征在分类识别中包含的差异信息以及信息的不同重要程度,而依据贡献度选择降维数据的特征维数,一方面能减小视图分割时引入噪声的概率,另一方面使得关键特征均匀划分到两个视图以更好地发挥协同作用。在其余3个数据集中,本文算法表现不是最佳,主要原因可能在于当初始有标记样本较少时,使用最小二乘法求解特征加权系数在该数据集上表现差,不能较好地拟合标记和样本的依赖关系,为后面的视图分割带来影响。尤其在数据集Abalone表现明显,但随着初始有标记样本数量的增加,上述情况会有很大改善,这点在实验2中得到了证明。

表2 有标记样本为10%,5个算法在9个数据集上两个视图的分类正确率

Tab.2 Labeled data is 10%, two views classification accuracy and standard deviation of 5 algorithms on 9 datasets %

数据集	CT		CTSMCKNN		CTDPC		CTALDPC		本文算法	
	h_1	h_2	h_1	h_2	h_1	h_2	h_1	h_2	h_1	h_2
Banknote Authentication	62.166 9	53.249 1	45.685 0	48.603 5	64.120 1	48.904 5	59.265 8	59.553 6	**75.592 7**	**77.372 8**
Abalone	50.413 8	50.983 5	50.375 4	51.141 6	**50.418 8**	51.022 3	50.418 5	**51.424 2**	32.597 0	34.599 0
Contraceptive	40.559 3	38.663 9	40.852 2	39.658 9	40.350 3	38.162 2	40.153 1	39.253 4	**42.549 3**	**42.521 2**
Wine	84.346 4	83.823 5	82.516 3	**85.065 4**	81.549 0	84.594 8	85.613 6	84.270 6	**86.039 4**	84.039 0
Liver Disorders	51.058 8	51.591 6	51.771 4	**52.882 4**	50.623 5	51.218 5	50.109 2	49.840 3	**53.684 0**	51.623 5
Eegeyestaage	48.440 6	48.863 8	48.144 2	48.210 9	49.173 6	49.432 6	**52.150 2**	**52.076 8**	50.060 0	49.166 2
Breast Cancer	93.231 3	91.630 0	92.921 4	91.759 5	92.008 3	91.190 6	91.554 0	91.663 0	92.047 7	**92.849 7**
Spectf Heart	47.447 3	46.868 9	59.549 9	59.549 9	56.472 9	62.008 5	35.247 3	36.029 7	**65.972 0**	**63.971 0**
Connectionist Bench	56.076 2	58.090 5	57.261 9	**60.638 1**	**59.209 5**	55.190 5	51.952 4	56.214 3	55.523 8	55.523 8

注:加粗表示算法在对应数据集上分类效果最好,下同

由表3可知,除数据集Eegeyestaage、Abalone和Connectionist Bench外,本文算法在其余6个数据集上分类正确率均高于对比算法。在数据集Eegeyestaage上本文算法的正确率略低于表现最好的CTALDPC算法,但优于其他3个对比算法。在数据集Abalone上,4个对比算法性能优于本文算法,而在数据集Connectionist Bench上,本文算法正确率仅高于CTALDPC,这是因为该数据集在有标记样本数量较少时不能通过有标记样本求得一个能较好地反映特征重要程度的加权系数,而加权系数会影响后续的降噪和视图分割过程,进而影响整个算法的分类性能。

表3 有标记样本比例为10%时,4个算法在9个数据集上的平均分类正确率

Tab.3 Labeled data is 10%, average classification accuracy of 5 algorithms on 9 datasets %

数据集	CT	CTSMCKNN	CTDPC	CTALDPC	本文算法
Banknote Authentication	57.708 0	47.144 3	56.512 3	59.409 7	**76.482 8**
Abalone	50.698 7	50.758 5	50.720 5	**50.921 3**	33.598 5
Contraceptive	39.611 6	40.255 5	39.256 3	39.703 3	**42.535 3**
Wine	84.085 0	83.790 8	83.071 9	84.942 1	**85.039 2**
Liver Disorders	51.325 2	52.326 9	50.921 0	49.974 8	**52.653 8**
Eegeyestaage	48.652 2	48.177 6	49.303 1	**52.113 5**	49.566 1
Breast Cancer	92.430 6	92.340 4	91.599 4	91.608 5	**92.448 7**
Spectf Heart	47.158 1	56.235 0	59.240 7	35.638 5	**64.971 5**
Connectionist Bench	57.083 3	**58.950 0**	57.200 0	54.083 3	55.523 8

2.3 实验 2

为了进一步探究本文算法在不同比例有标记样本下的性能表现，实验 2 将初始有标记样本比例设定为 10%，20%，30%，40%和 50%，最终实验结果见图 3。

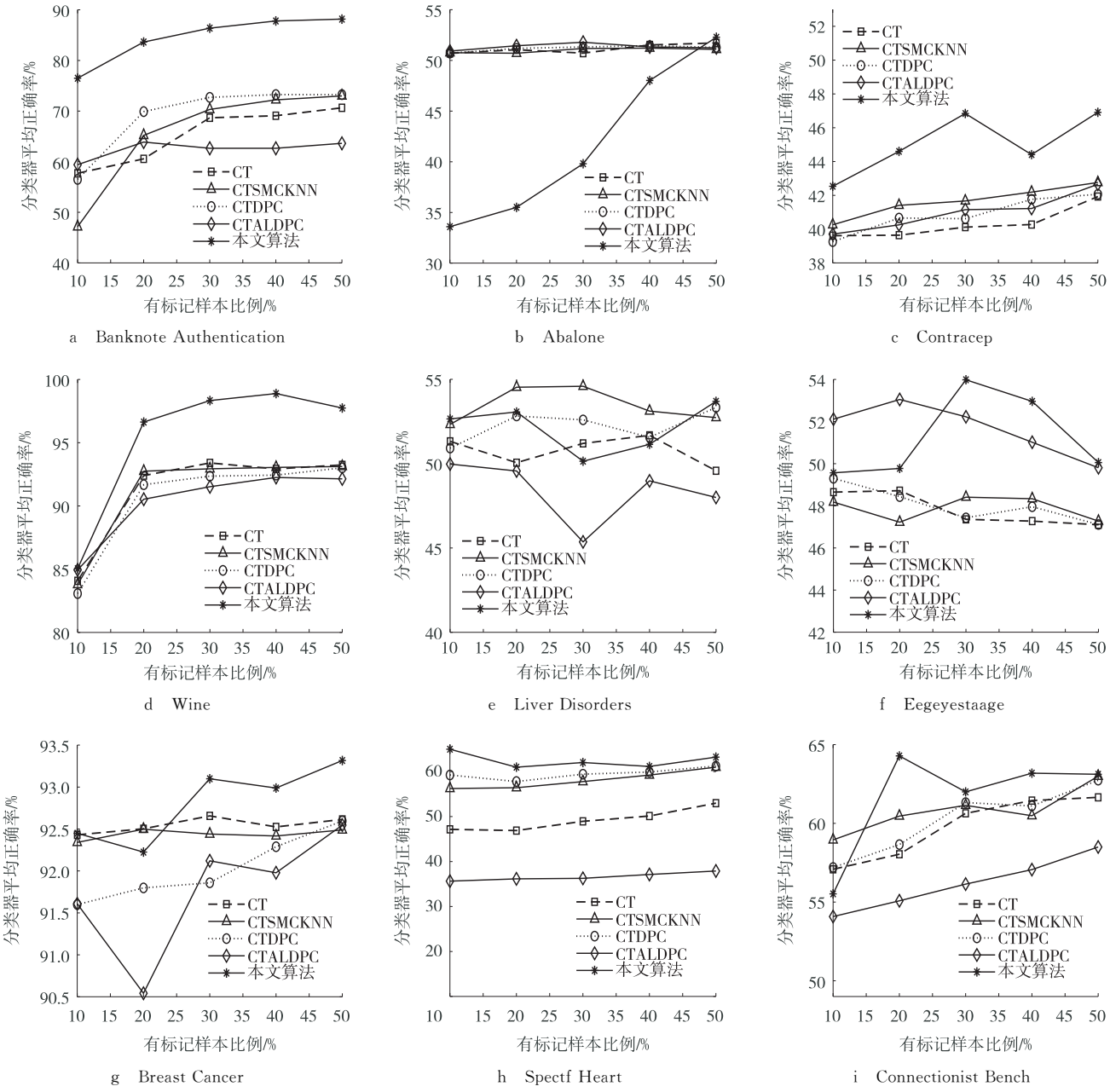


图 3　有标记样本数量与平均分类正确率的关系曲线图

Fig. 3　Relation curve between the number of label samples and the average classification accuracy

从图 3 可以看出，随着初始有标记样本比列增加，算法在 9 个数据集上的分类正确率在逐步提升，这是由于在训练初期包含更多的有标记样本能训练出一个分类能力较强的初始分类器，有效避免初期产生的误差在迭代中累积出错。在数据集 Banknote Authentication，Contraceptive，Wine 和 Spectf Heart 中，本文算法在有标记样本为 10%，20%，30%，40%和 50%时性能表现优于所有对比算法。主要原因一方面在于使用 WPCA 对数据预处理能有效过滤掉数据中噪声并使关键特征在分类识别过程中占有更大比重；另一方面则是基于“双拐点”聚类中心选择法的密度峰值聚类在数据集上表现较好，能正确修正标记不一致样本，减少被误分类的概率。由图 3f，g 可知，在数据集 Eegeyestaage 和 Breast Cancer 中，当有标记样本为 10%和 20%时，本文算法正确率略低于 CTSMCKNN 和 CTALDPC，但随着有标记样本数量的增加，本文算法正确率高于全部对比算法。这是因为在有标记样本较少时 CTSMCKNN 算法能利用半监督聚类能选择出较好表现数据空间结构的样本，而 CTALDPC

则利用主动学习选择模糊度高的样本改善分类器。在数据集 Connectionist Bench 中,当有标记样本为 20%时,本文算法的分类性得到极大提升并高于对比算法,而随后当有标记样本进一步增加,性能则逐渐趋于稳定。从图 3b 可以看到,在数据集 Abalone 上随着标记样本增加,本文算法性能快速增长并且在 50%时高于所有对比算法。由于该数据集分布极不规则且噪声点较多,所以密度峰值聚类对样本的修正效果差,可能会误标记一些样本。但是随着有标记样本的增加,这种影响会逐渐降低,因为 W-PCA 能根据大量有标记样本得到一个较好的特征加权系数,从而在降维过程中更好的实现噪声过滤。

2.4 实验 3

本次实验为了探究有标记样本比例为 10%时,算法在样本修正上的表现。在 4 个对比算法中,只有算法 CTSMCKNN 和 CTDPC 对两个视图分类器标记不一致样本进行了修正,所以将本文算法同上述两个算法进行对比实验,最后结果如表 4 所示。其中"+"代表正确修正,"−"错误修正。

从表 4 可以看出,本文算法在 7 个数据集上的修正率优于对比算法。由于聚类在数据集 Abalone 上表现不佳,而 CTDPC 和本文算法都是基于密度峰值聚类来对样本修正,所以导致二者都出现了错误修正情况。在数据集 Liver Disorders 上本文算法表现较差,可能由于基于少量有标记样本不能取得一个较好的特征加权系数造成。表 4 说明本文算法在大部分数据集上都能取得一个较好的修正效果,证明了本文算法中对样本修正方法的有效性。

表 4 3 个算法在 9 个数据集上的修正率

Tab. 4 Correction rate of 3 algorithms on 9 data sets %

数据集	CTSMCKNN	CTDPC	本文算法	数据集	CTSMCKNN	CTDPC	本文算法
Banknote Authentication	+1.640 9	+4.030 4	**+17.704 2**	Eegeyestaage	+1.053 5	+0.074 8	**+6.626 9**
Abalone	**+0.460 0**	−0.609 9	−10.907 0	Breast cancer	−0.216 4	−0.797 8	**+0.442 7**
contraceptive	+1.549 0	+0.500 2	**+6.722 9**	Spectf Heart	+8.831 9	+10.482 9	**+13.079 8**
wine	−0.464 1	−1.042 5	**+4.931 3**	Connectionist bench	+0.509 5	+0.683 3	**+1.361 9**
Liver Disorders	**+3.387 4**	+1.396 7	−1.055 4				

2.5 实验 4

实验 4 主要对算法的运行时间进行分析,采用 10 次十折交叉验证的平均时间作为算法的运行时间,图 4 中给出了 5 个算法在 9 个数据集上运行的总时间。

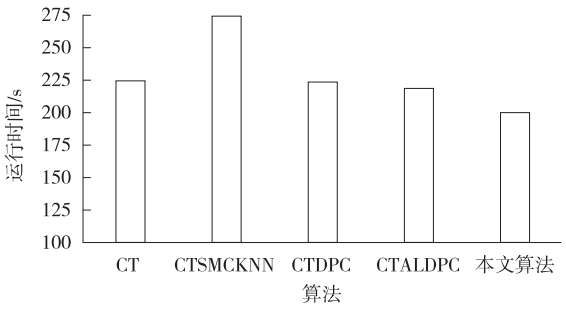


图 4 5 个算法在 9 个数据集上的总运行时间

Fig. 4 Total running time of 5 algorithms on 9 datasets

从图 4 可以看出,本文算法运行时间最短,其次是 CTALDPC 算法,最长是 CTSMCKNN 算法。主要原因是本文算法中基于 WPCA 对数据的预处理,能有效实现数据降维和噪声过滤,高维数据降维后能极大地减少计算量从而降低算法的运行时间,而过滤噪声能避免引入误差进而减少算法修正样本的时间。在 CTSMCKNN 算法中,由于 KNN 修正样本时会耗费大量时间计算样本间距离,尤其当数据量较大时运行时间会快速增加,所以导致算法运行速度慢。

3 结论

针对协同训练算法中视图分割时噪声引入和无标记样本标记不一致带来错误累积的问题,本文提出了一种基于加权主成分分析和改进密度峰值聚类的协同训练算法。该算法首先利用 WPCA 对数据预处理,在加权变换后的数据上进行降维,然后选取贡献度高的特征匀划分到两个视图达到过滤噪声的目的,最后对标记不一致的样本采用密度峰值聚类修正样本类别,减少分类过程中错误的累积和样本误分类的概率。通过 4 组实验证明,本文算法的分类正确率高,运行效率更高。在后续工作中,将进一步研究针对不同数据集,如何自适应求得

特征加权系数和不存在两个充分冗余视图时如何有效地进行视图分割。

参考文献：

[1] NAMRATA P, SOMNATH R, SUSHMITA P, et al. Genome-wide analysis of multi-view data of miRNA-seq to identify miRNA biomarkers for stomach cancer[J]. Journal of Biomedical Informatics, 2019, 97: 103254.

[2] LI J X, ZHANG B, LU G M, et al. Generative multi-view and multi-feature learning for classification[J]. Information Fusion, 2019, 45(1): 215-226.

[3] 龚彦鹭，吕佳. 结合半监督聚类和加权 KNN 的协同训练方法[J]. 计算机工程与应用，2019，55(22)：114-118.
GONG Y L, LÜ J. Co-training method combined with semi-supervised clustering and weighted *K*-nearest neighbor[J]. Computer Engineering and Applications, 2019, 55(22): 114-118.

[4] HASSANI M S, GREEN J R. Multi-view co-training for microRNA prediction[J]. Scientific Reports, 2019, 9(1): 1093-1103.

[5] LIU A A, XU N, NIE W Z, et al. Benchmarking a multimodal and multiview and interactive dataset for human action recognition [J]. IEEE Transactions on Cybernetics, 2017, 47(7): 1781-1794.

[6] TANG J J, TIAN Y J, LIU X H, et al. Improved multi-view privileged support vector machine[J]. Neural Networks, 2018, 106 (16): 96-109.

[7] WEN J, XU Y, LIU H. Incomplete multiview spectral clustering with adaptive graph learning[J]. IEEE Transactions on Cybernetics, 2020, 50(4): 1418-1429.

[8] GUPTA S, GUPTA A. Dealing with noise problem in machine learning data-sets: a systematic review[J]. Procedia Computer Science, 2019, 161(11): 466-474.

[9] LIANG Y, LU S, WENG R, et al. Unsupervised noise-robust feature extraction for aerial image classification[J]. Science China Technological Sciences, 2020, 63: 1406-1415.

[10] ZHOU T, ZHANG C Q, GONG C, et al. Multiview latent space learning with feature redundancy minimization[J]. IEEE Transactions on Cybernetics, 2020, 50(4): 1655-1668.

[11] ZHAO F, LI J, ZHANG L, et al. Multi-view face recognition using deep neural networks[J]. Future Generation Computer Systems, 2020, 111(12): 375-380 .

[12] ZHANG H, JIANG L X, YU L J. Class-specific attribute value weighting for naive Bayes[J]. Information Sciences, 2020, 508 (2): 260-274.

[13] ZHONG J, WANG N, LIN Q, et al. Weighted feature selection via discriminative sparse multi-view learning[J]. Knowledge-Based Systems, 2019, 178(C): 132-148.

[14] 王永欣，张化祥，王爽. 基于属性加权的主成分分析算法[J]. 济南大学学报(自然科学版)，2015，29(6)：438-443.
WANG Y X, ZAHNG H X, WANG S. Anattribute-weighted principal-component analysis algorithm[J]. Journal of University of Jinan (Science and Technology), 2015, 29(6): 438-443.

[15] TIAN W D, ZHANG G X, ZHANG X, et al. PCA weight and Johnson transformation based alarm threshold optimization in chemical processes[J]. Chinese Journal of Chemical Engineering, 2018, 26(8): 1653-1661.

[16] 黄欣研，李玲，辛云宏. WPCA-LDA：一种数据分类新方法[J]. 计算机应用研究，2017，34(6)：1650-1653.
HUANG X Y, LI L, XIN Y H. WPCA-LDA: new method of data classification[J]. Application Research of Computers, 2017, 34 (6): 1650-1653.

[17] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1497.

[18] 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.

[19] 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.
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.

[20] XU X, DING S F, WANG L J, et al. A robust density peaks clustering algorithm with density-sensitive similarity[J]. Knowledge-Based Systems, 2020, 200(14): 832-836.

[21] WANG F, ZHOU J Y, TIAN Y, et al. Intradialytic blood pressure pattern recognition based on density peak clustering[J]. Journal of Biomedical Informatics, 2018, 83(18): 33-39.

[22] SIERANOJA S, FRÄNTI P. Fast and general density peaks clustering[J]. Pattern Recognition Letters, 2019, 128(10): 551-

558.

[23] 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, LÜ J. Co-training algorithm with combination of active learning and density peak clustering[J]. Journal of Computer Applications,2019,39(8):2297-2301.

A Co-Training Agorithm Based on WPCA and Improved Density Clustering

GONG Xu, LÜ Jia

(Chongqing Center of Engineering Technology Research on Digital Agriculture Service, College of Computer and Information Sciences, Chongqing Normal University, Chongqing Normal University, Chongqing 401331, China)

Abstract: [**Purposes**]In the co-training algorithm, the noise effect is not considered in view segmentation and inconsistent labeling of unlabeled samples by two view classifiers. Aimed at the above questions, a co-training algorithm based on weighted principal component analysis (WPCA) and improved density peak clustering is proposed. [**Methods**]Firstly, the WPCA is introduced into data preprocessing. The weighted coefficient is obtained by linear fitting the dependency between data and the class in initial labeled samples. Then, the dimension of weighted transformed data is reduced and high contribution features are extracted for view segmentation. This strategy can filter the noise in view segmentation and key features are evenly divided into two views, so it can better achieve the synergy of the two classifiers. At the same time, a "double turning point" method is proposed to automatically select the cluster center in the density peak clustering. Then, the improved density peak clustering is utilized to re-classify the samples of inconsistent label, which can decline the probability of sample misclassification. [**Findings**] Compared with the comparison algorithm, the proposed algorithm has better classification accuracy and algorithm stability. [**Conclusions**] Four experiments on 9 UCI datasets show that the proposed algorithm has a great improvement in classification accuracy and efficiency.

Keywords: co-training; weighted PCA; density peak clustering; "double turning point" method

(责任编辑 黄 颖)