软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(2):539−554 [doi: 10.13328/j.cnki.jos.006139]

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

考虑标记间协作的标记分布学习*

李睿钰, 祝继华, 刘新媛

(西安交通大学 软件学院, 陕西 西安 710049)

通信作者: 祝继华, E-mail: zhujh@xjtu.edu.cn

摘　要: 近些年来, 作为一种新的有监督学习范式, 标记分布学习(LDL)已被应用到多个领域, 如人脸年龄估计、头部姿态估计、电影评分预测、公共视频监控中的人群计数等, 并且在这些领域的相关任务上取得了一定性能上的进展. 最近几年, 很多关于标记分布学习的算法在解决标记分布学习问题时考虑到了标记之间的相关性, 但是现有方法大多将标记相关性作为先验知识, 这可能无法正确刻画标记之间的真实关系. 此外, 标记相关性通常用于在训练阶段调整假设空间, 而最终的标记预测并未显式利用标记间的相关性. 因此, 提出一种新的标记分布学习方法——考虑标记间协作的标记分布学习(LDLCL). 该方法旨在训练期望模型的同时, 显式地考虑标记间的相关预测. 具体来讲, 首先提出假设: 对于每个标记, 最终的预测结果涉及到它自己的预测和其他标记的预测之间的协作. 基于这一假设, 提出一种通过标记空间中的稀疏重构来学习标记相关性的新方法; 然后, 将学习到的标记相关性无缝集成到模型训练中; 最终, 在标记预测时使用学习到的标记相关性. 大量的实验结果表明, 该方法优于近期的同类方法.

关键词: 标记分布学习; 标记相关性; 样本相似性; 标记分布

中图法分类号: TP18

中文引用格式: 李睿钰, 祝继华, 刘新媛. 考虑标记间协作的标记分布学习. 软件学报, 2022, 33(2): 539–554. http://www.jos.org.cn/1000-9825/6139.htm

英文引用格式: Li RY, Zhu JH, Liu XY. Label Distribution Learning with Collaboration among Labels. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 539−554 (in Chinese). http://www.jos.org.cn/1000-9825/6139.htm

Label Distribution Learning with Collaboration among Labels

LI Rui-Yu, ZHU Ji-Hua, LIU Xin-Yuan

(School of Software Engineering, Xi’an Jiaotong University, Xi’an 710049, China)

Abstract: In last few years, as a new supervised learning paradigm, label distribution learning (LDL) has been applied to many fields and shown good results in these fields, such as face age estimation, head posture estimation, movie score prediction and crowd count in public video surveillance. Recently, the correlations between labels have been considered in some algorithms when solving the problem of label distribution learning. However, most of the existing methods take label correlations as a prior knowledge, which may not be able to correctly describe the real relationship between labels. In addition, label correlations are usually used to regularize the hypothesis space in the training phase, while the final label distribution prediction does not use these correlations explicitly. Therefore, this study proposes a new label distribution learning method, label distribution learning with collaboration among labels (LDLCL), which aims to explicitly consider the correlated predictions of labels while training the expected model. Specifically, the hypothesis is first proposed: for each label, the final prediction involves the cooperation between its own prediction and other labels’ predictions. Based on this assumption, a new method is proposed to learn label correlations by sparse reconstruction in label space. Then, the learned label correlations are seamlessly integrated into model training, and finally the learned label correlations are used in label prediction. Sufficient experimental results show that the proposedapproach is superior to other similar methods.

* 基金项目: 江苏省自然科学基金(BK20191287)

收稿时间: 2020-06-09; 修改时间: 2020-07-16, 2020-08-09; 采用时间: 2020-08-20

Key words: label distribution learning; label correlations; instance similarity; label distribution

多义性学习在机器学习研究中一直是一个热门的领域. 单标记学习(single-label learning, SLL)和多标记学习(multi-label learning, MLL)[1,2]是目前解决标记多义性问题(label ambiguity problem)的两种成熟的学习范式: 前者假设每个实例只能与一个预定义的标记相关联, 而后者假设每个实例可能具有一组类标记. 大量研究[3−5]表明: 多标记学习是一种应用更广泛的学习范式, 能够解决更复杂的标记多义性问题. 然而, 单标记学习和多标记学习都无法回答更进一步的问题, 即标记以多大的程度描述样本. 现实世界中, 数据的标记往往带有对样本的描述程度, 比如歌曲评分预测, 由于对歌曲的喜爱程度评分由多个个体进行标注, 不同的分数表达了评分者对歌曲的不同感受, 对这些分数取平均得到最终评分并不恰当, 因为歌曲在不同人群中的评分可能存在两极分化, 比如青少年和老年人对于嘻哈歌曲的不同态度, 平均得分不能全面反映人们对于歌曲的真实评价. 为了解决这一问题, Geng 等人[6]提出了一种新的多义性学习范式——标记分布学习(LDL), 正式给出了标记分布学习的定义, 总结了标记分布学习在处理标记多义性问题上的优势. 作为多标记学习的自然延伸, 标记分布学习可以更加广泛地应用到更多解决标记多义性问题的场景中.

假定有 n 个样本, 每个样本由一个 d 维特征向量表示, 则所有样本组成样本集 $\boldsymbol{X}=[\boldsymbol{x}_1;\boldsymbol{x}_2;\ldots;\boldsymbol{x}_n]\in R^{n\times d}$. 使用 l 个标记对每个样本进行标注, 所有标记组成有限标记集 $\boldsymbol{Y}=\{y_1,y_2,\ldots,y_l\}$. 使用描述度 $d_{\boldsymbol{x}}^{y}$($d_{\boldsymbol{x}}^{y}\in[0,1]$且$\sum_y d_{\boldsymbol{x}}^{y}=1$)来量化表示标记 y 对样本 $\boldsymbol{x}_i$ 的描述程度, 则所有标记对样本的描述度组成样本的标记分布 $\boldsymbol{D}_i=[d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2},\ldots,d_{\boldsymbol{x}_i}^{y_l}]$, 所有样本的标记分布组成样本集对应的标记分布空间 $\boldsymbol{D}=[\boldsymbol{D}_1;\boldsymbol{D}_2;\ldots;\boldsymbol{D}_n]\in R^{n\times l}$.

给定一个训练集 $T=\{(\boldsymbol{x}_1,\boldsymbol{D}_1),(\boldsymbol{x}_2,\boldsymbol{D}_2),\ldots,(\boldsymbol{x}_n,\boldsymbol{D}_n)\}$, 其中, $\boldsymbol{x}_i\in\boldsymbol{X}$ 表示一个样本, $\boldsymbol{D}_i\in\boldsymbol{D}$ 为样本 $\boldsymbol{x}_i$ 对应的标记分布, 寻找一个从输入特征空间 $\boldsymbol{X}$ 到标记分布空间 $\boldsymbol{D}$ 的映射的过程称为标记分布学习. 假设 $\boldsymbol{X}$ 经过映射得到的最终标记分布预测空间为 $\hat{\boldsymbol{D}}$, 则标记分布学习的目标是使 $\hat{\boldsymbol{D}}$ 和 $\boldsymbol{D}$ 尽可能相似.

由于标记分布可以看作是一种标记关系, 因此在学习过程中, 不应该忽略标记之间的相关性. 例如在图像注释中, 假设沙漠和太阳是强相关的, 则当一张图像与沙漠强相关时, 图像有很大可能与太阳相关. 基于此, 近年来, 一些关于标记分布学习的研究考虑了标记间的相关性. Wang 等人[7]用二进制编码样本后使用调整余弦相似度(the adjusted cosine similarity, ACS)来计算标记相关性. 在文本情感识别领域, Zhou 等人[8]使用普鲁契克情感色轮(Plutchik's wheel of motions)[9]得到基于情感先验信息的标记相关性. Zhou 等人[10]、Jia 等人[11]应用 Pearson 相关系数计算标记相关性. Ren 等人[12]通过低秩约束获得标记间的相关性, 提出了 LDL-LCLR 算法. 在 Zhang 等人[13]提出的 COS-LDL 算法中, 使用基于余弦的距离映射来度量标记相关性. Zhao 等人[14]通过将 LDL 问题转换为最优传输(optimal transport, OT)问题, 将 OT 问题中的 ground metric 用于刻画标记相关性. 然而, 以上研究中, 标记相关性大都只是通过常用的相似性度量来量化标记两两之间的相似性, 可能无法反映标记之间的复杂关系; 此外, 这些方法只在训练过程中利用标记相关性, 在最终的标记预测中并没有显式利用此相关性. 但在预测某一标记的分布时, 考虑此标记和其他标记间的相关关系是很有必要的, 因为采用一些标记描述某一个体时, 所有标记都不是互不相关的, 标记空间中每个标记的状态都会对整个标记分布产生影响. 标记相关性信息在训练阶段辅助模型训练, 在标记分布预测时也能够辅助模型预测. 直接使用训练学习到的模型参数将测试样本从样本空间(原始样本空间或经过空间变换的潜在样本空间)映射到标记分布空间, 只是单纯地将标记看作样本空间特征的线性组合, 标记之间各自独立. 但是标记间的相关信息是我们获取到的关于标记的有利信息, 利用此信息可以辅助预测. 例如: 在图 1 所示的图像注释任务中, 我们收集到了很多风景图, 选择标记瀑布(y_1)、岩石(y_2)、树木(y_3)、青苔(y_4)对这些图像进行描述, 所有图像与其对应的标记分布组成了图像注释训练集. 假定利用此训练集我们获取到了标记间的相关关系矩阵 $\boldsymbol{G}=[\boldsymbol{g}_1,\boldsymbol{g}_2,\boldsymbol{g}_3,\boldsymbol{g}_4]$和特征映射矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3,\boldsymbol{w}_4]$, 其中, 列向量 $\boldsymbol{g}_i$(i=1,2,3,4)表示每个标记与标记 y_i 的相关性, 列向量 $\boldsymbol{w}_i$(i=1,2,3,4)为求样本描述度 $d_{\boldsymbol{x}}^{y_i}$ 时样本特征向量 $\boldsymbol{x}$ 对应的特征映射向量. 对于测试样本, 图像中存在大面积的岩石、部分青苔、一条瀑布和一小棵树, 由于树在整个图像中所占比例很小且整张图片色调偏暗, 图像特征向量 $\boldsymbol{x}_{test}$ 中与 y_3 相关的特征值会很小, 在不考虑标记间的相关关系时, 直接对样本 $\boldsymbol{x}_{test}$ 经过特征映射 $\boldsymbol{W}$ 求标

记分布，求得的 y_3 对应的描述度很小甚至为 0，如图 1 中 $\boldsymbol{d}_{before}$ 所示. 若在预测时考虑标记间的相关信息 $\boldsymbol{G}$，则得到调整后的标记分布为 $\boldsymbol{d}_{after}$. 可以看出：由于考虑了每个标记对标记 y_3 的影响，使得 y_3 对应的描述度由 0 上升至 0.18；同样地，其他标记的描述度也因为考虑了标记间的相关关系发生了改变. 调整后的标记分布 $\boldsymbol{d}_{after}$ 更全面、细致地描述了图片内容，使标记分布预测更加准确并且贴合真实情况.

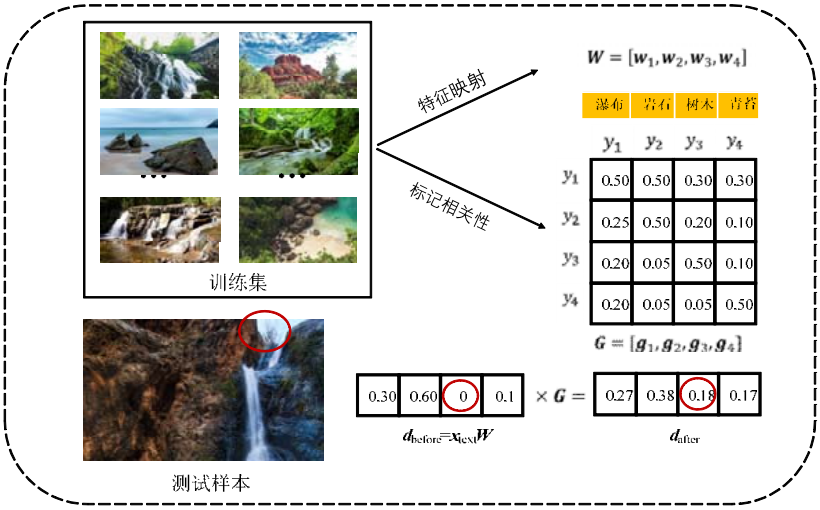


图 1　预测时考虑标记相关性作用的示意图

为了解决上述问题，我们提出了一种新的关键假设：对于每一个独立的标记，它的最终预测涉及到其自身的预测和其他标记的预测之间的协作. 基于此假设，我们提出了一个新的标记学习方法——考虑标记间协作的标记分布学习(label distribution learning with collaboration among labels, LDLCL). 我们首先通过标记空间中的稀疏重构来学习标记相关性矩阵，得到相关性矩阵后，在训练和预测阶段均使用学习到的标记相关性矩阵实现标记间的协作. 具体来讲，我们在训练阶段引入了标记独立嵌入作为潜在的标记分布空间，目的是在训练模型参数的同时，用学习到的标记相关性拟合最终预测. 我们在 14 个标记分布学习中广泛使用的数据集上对 LDLCL 算法的有效性进行了验证，实验结果表明，我们的方法优于同类方法.

本文主要贡献如下:

(1) 提出在计算标记相关性时考虑标记间的协同作用来获得标记相关性矩阵;

(2) 使用核函数度量样本间的相似性，同时，在计算相似性时考虑了样本标记个数的影响. 在训练过程中考虑了训练样本间的相似性，在标记预测时考虑了测试样本与训练样本间的相似性;

(3) 在标记分布学习模型训练和样本标记预测两个阶段都显式使用了标记相关性.

1　相关工作

1.1　标记分布学习

标记分布学习范式提出后，已被广泛应用到各种解决标记多义性问题的研究中. 为了解决年龄估计问题，Geng 等人[15]首次提出了 IIS-LDL 方法. 基于该方法，一张人脸图像可以在训练过程中不但提供其对应的确切年龄，同时还能够提供与确切年龄相近的年龄区间，从而有效提高了样本包含的信息量，进而提高了年龄估计的准确度. 为了进一步提高年龄估计的准确性，文献[16]提出了基于条件概率的神经网络算法 CPNN. 该方法构造了一个简单的 3 层网络，将样本特征和目标年龄同时作为输入变量，利用网络隐藏层自动选择特征完成年龄估计. 在文献[16]首次使用神经网络后，Yang 等人[17]通过将标记分布学习和深度学习相结合，提出了深度标记分布学习(deep label distribution learning, DLDL)算法. Gao 等人[18]对 DLDL 算法进行了改进，使深度标

记分布学习不仅可以应用到年龄估计, 还可以用于头部姿态估计, 并且能够提高多标记分类和语义分割任务的识别性能. 针对标记分布学习训练过程计算量大、时间复杂度高的问题, Wang 等人[7]提出的 BC-LDL 算法使用定长二进制串编码样本, 通过计算二进制码间的海明(Hamming)距离, 选择最近邻 k 个样本的二进制码生成测试样本的标记分布. 之后, 为了解决 BC-LDL 算法中采用二进制编码样本带来的量化误差和长二进制码导致的模型训练耗时问题, Wang 等人[19]设计了一个有效的离散二进制编码框架对样本编码. 该框架集成了样本间语义相似度信息和原始标记分布用于学习具有高度区分性的二进制码, 形成了 DBC-LDL 算法. Wang 等人[20]提出的 KELM-LDL 算法通过高斯核函数将特征映射到高维空间, 然后对原标记空间建立核极限学习机回归模型求得输出权值, 以此减少计算量. 与大多数 LDL 算法中直接使用样本的全部特征进行模型训练不同, 为了减少冗余和不相关特征对算法性能的影响, Ren 等人[21]提出对样本特征进行选择, 通过对样本特征映射权重矩阵 W 进行 $L1$ 范数约束选择标记特定属性(label-specific features), 同时, 用 $L2$ 范数约束特征映射权重矩阵 M 选择共有特征, 使用选择到的特征完成 LDL 模型训练和样本标记分布预测. Xu 等人[22]提出的 LSE-LDL 算法直接将样本空间映射到一个潜在语义空间中, 使模型在学习标记分布映射的同时进行特征选择. 此外, Zhao 等人[23]对标记分布学习中选择什么样的距离度量作为目标函数进行了实验, 并基于实验结果提出了一些建议.

现有的 LDL 算法可以分为 3 类: 基于问题转换(problem transformation, PT)策略的算法、算法自适应(algorithm adaption, AA)方法和专门为解决 LDL 问题设计的专用算法(specialized algorithm, SA). 问题转换算法将 LDL 问题转换为多个 SLL 问题, 然后利用现有的 SLL 方法来解决. 例如, 将 SVM 转换为 PT-SVM 和将 Bayes 转换为 PT-Bayes. 算法自适应的主要思想是: 将传统算法修改为适用于 LDL 的学习范式, 如 AA-KNN 算法[6]和 LogitBoost 算法[24]. 近些年来, 专门针对 LDL 问题设计的算法通过直接模拟每个标记对特定实例的相对重要性, 取得了很好的效果.

1.2 标记分布学习中的标记相关性

近几年内, 陆续有文献针对 LDL 中标记之间的相关性进行了研究. Jia 等人[11]提出的 LDLLC 算法中, 使用每个标记对应的特征映射权重矩阵 W 中列向量之间的欧式距离表示标记两两间的相似性, 同时使用 Pearson 相关系数约束标记之间的相似度. Ren 等人[12]提出的 LDL-LCLR 算法同时利用标记之间的全局相关性和局部相关性为训练模型提供更多的信息. 具体来说, 基于低秩近似的标记相关性矩阵被应用于捕获全局标记相关性, 同时采用局部样本间的标记相关性来修正标记相关性矩阵. Jia 等人[25]提出的 LDL-SCL 算法考虑了标记间的局部相关性, 为了在局部样本上使用标记相关性, 对局部样本的影响进行编码, 并基于不同的样本集群设计一个局部相关向量作为每个样本的附加特征, 通过同时利用样本的原始特征和附加特征来预测样本的标记分布. 以上关于 LDL 的研究都考虑了标记相关性且取得了不错的效果, 但都只是关注了标记两两之间的相关性, 并且只在训练过程中利用了标记相关性. 受到多标记学习领域 Feng 等人[26]考虑标记间协作关系的启发, 在标记分布学习问题中, 我们提出考虑每个标记与其他标记间的协作关系来完成标记分布学习任务.

2 考虑标记间协作的标记分布学习

首先对下文中要使用的符号进行说明. 为了与相关工作部分符号统一, 我们使用 $\boldsymbol{X}=[\boldsymbol{x}_1;\boldsymbol{x}_2;\ldots;\boldsymbol{x}_n]\in R^{n\times d}$ 表示输入特征空间, $\boldsymbol{Y}=\{y_1,y_2,\ldots,y_l\}$表示有限标记集, $\boldsymbol{D}=[\boldsymbol{D}_1;\boldsymbol{D}_2;\ldots;\boldsymbol{D}_n]\in R^{n\times l}$ 表示输入空间对应的标记分布空间. 其中, n 为样本个数, d 为样本的特征维数, l 为标记空间中包含的标记个数. 除此之外, 我们使用 $\boldsymbol{D}^j\in R^n$ 表示矩阵 $\boldsymbol{D}$ 的第 j 列向量并称 $\boldsymbol{D}^j$ 为一个标记向量, 标记向量 $\boldsymbol{D}^j$ 中每个元素 $\boldsymbol{D}_i^j(i=1,2,...,n)$ 表示标记 y 对样本 $\boldsymbol{x}_i$ 的描述度, $\boldsymbol{D}^{-j}=[\boldsymbol{D}^1,\boldsymbol{D}^2,\ldots,\boldsymbol{D}^{j-1},\boldsymbol{D}^{j+1},\ldots,\boldsymbol{D}^l]\in R^{n\times(l-1)}$表示标记分布空间 $\boldsymbol{D}$ 除标记向量 $\boldsymbol{D}^j$ 外剩余标记向量组成的矩阵.

2.1 标记相关性矩阵

在 LDLCL 方法中, 我们首先需要学习一个标记相关性矩阵 $\boldsymbol{S}^*=[\boldsymbol{S}_1^*,\boldsymbol{S}_2^*,...,\boldsymbol{S}_l^*]\in R^{l\times l}$ 用来描述标记之间的

协作关系，$\boldsymbol{S}^*$中的元素 s_{ij}^* 反映第 i 个标记对第 j 个标记的贡献程度，对角线元素 $s_{ii}^*=(1-\alpha)/\alpha=1/\alpha-1$ $(i=1,2,...,l)$ 表示第 i 个标记对自身的贡献程度. 在假设每个标记的最终预测涉及到其自身预测和其他标记预测之间的协作的前提下，我们将给定的标记矩阵 $\boldsymbol{D}$ 作为最终预测，并通过以下方式学习标记相关矩阵 $\boldsymbol{S}^*$:

$$\min_{S^*}\frac{1}{2}\|\alpha\boldsymbol{D}\boldsymbol{S}^*-\boldsymbol{D}\|_F^2 \tag{1}$$

其中，$\alpha\in[0,1]$为折衷参数，用来控制每个标记与其他标记间协作的程度. 对公式(1)中$\alpha\boldsymbol{D}\boldsymbol{S}^*$做因式分解，同时引入新的标记相关性矩阵 $\boldsymbol{S}=\boldsymbol{S}^*-diag(\boldsymbol{S}^*)$，得到公式(2):

$$\alpha\boldsymbol{D}\boldsymbol{S}^*=\alpha\boldsymbol{D}(\boldsymbol{S}+(1/\alpha-1)\boldsymbol{I})=(1-\alpha)\boldsymbol{D}+\alpha\boldsymbol{D}\boldsymbol{S} \tag{2}$$

其中，$\boldsymbol{S}$ 为不关注每个标记自身，只考虑其他标记对此标记的贡献程度的标记相关性矩阵. 公式(2)表示：每个标记可以由其自身和其他标记共同协作表示，两者协作时各自的贡献程度分别为$(1-\alpha)$和α. 将公式(2)代入公式(1)，得到关于标记相关性矩阵 $\boldsymbol{S}$ 的优化问题:

$$\min_{S}\frac{1}{2}\|\alpha\boldsymbol{D}(\boldsymbol{S}-\boldsymbol{I})\|_F^2 \tag{3}$$

由于矩阵 $\boldsymbol{S}$ 中的对角线元素固定为 0，所以需要去掉对角线元素求解 $\boldsymbol{S}$. 对每个标记，使用列向量 $\boldsymbol{s}_j\in R^{l-1}(j=1,2,\ldots,l)$来表示其他$(l-1)$个标记对此标记的贡献程度，对每个标记分别求解其对应的 $\boldsymbol{s}_j$，可将公式(3)改写为如下公式(4):

$$\min_{s_j}\frac{1}{2}\|\alpha\boldsymbol{D}^{-j}\boldsymbol{s}_j-\alpha\boldsymbol{D}^j\|_2^2 \tag{4}$$

考虑到每个标记只会与其他标记中的几个标记有关，而不会与大部分标记都相关，因此，标记之间的协作关系应该满足稀疏约束. 加入稀疏约束后，公式(4)可扩展为公式(5):

$$\min_{s_j}\frac{\alpha^2}{2}\|\boldsymbol{D}^{-j}\boldsymbol{s}_j-\boldsymbol{D}^j\|_2^2+\hat{\lambda}\|\boldsymbol{s}_j\|_1 \tag{5}$$

其中，$\hat{\lambda}$ 为权衡参数，控制 $\boldsymbol{s}_j$ 的稀疏程度. 为了清晰地说明标记间的协作关系，我们在图 2 中给出了标记相关性示例. 图 1 中假设样本集中共有 5 个样本，每个样本使用 6 个标记描述，组成标记分布集合 $\boldsymbol{D}\in R^{5\times6}$，那么每个标记向量$\{\boldsymbol{D}^1,\boldsymbol{D}^2,\ldots,\boldsymbol{D}^6\}$都可以由此标记向量本身和其他标记向量集合两个因素协作表示. 图 2 中，不同色块组成的矩阵为构成此矩阵的数值元素可视化的结果，数值元素范围为[0,1]，对应颜色范围由红过渡到蓝(0 对应深蓝色，1 对应正红色). 从图 2 中可以看出：标记相关性矩阵 $\boldsymbol{S}$ 中的对角元素均为深蓝色，表示对角线元素均为 0，同时，$\boldsymbol{S}$ 中相同或相近颜色的色块较多，表明 $\boldsymbol{S}$ 为稀疏矩阵.

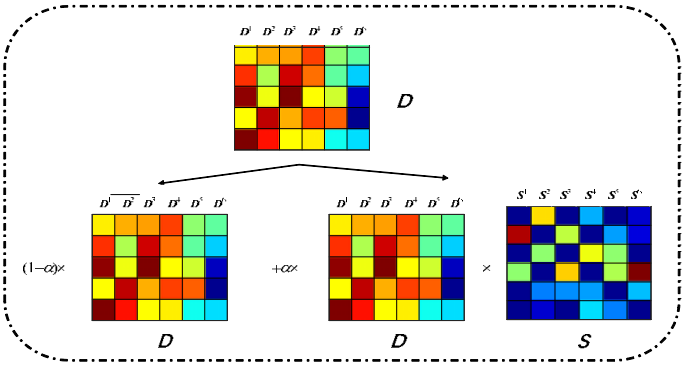


图 2　基于标记间协作的标记相关性示意图，其中，$s_{ii}=0(i=1,2,\ldots,6)$

公式(5)中，未知变量 $\boldsymbol{s}_j$ 单独存在 $L1$ 范数(L1-norm)约束，同时以$(\boldsymbol{D}^{-j}\boldsymbol{s}_j-\boldsymbol{D}^j)$的形式受 $L2$ 范数约束，直接求解 $\boldsymbol{s}_j$ 比较困难. 为了方便求解，参照范数相关问题的求解方法，我们引入新的变量 $\boldsymbol{z}_j$，同时令 $\lambda=\hat{\lambda}/\alpha^2$，将公式(5)转化为公式(6):

$$\left.\begin{aligned}&\min_{\boldsymbol{s}_j,\boldsymbol{z}_j}\frac{1}{2}\|\boldsymbol{D}^{-j}\boldsymbol{s}_j-\boldsymbol{D}^j\|_2^2+\lambda\|\boldsymbol{z}_j\|_1\\&\text{s.t. }\boldsymbol{s}_j-\boldsymbol{z}_j=0\end{aligned}\right\}\tag{6}$$

公式(6)为含有两个变量的等式约束优化问题, 可以使用交替方向乘子法(alternating direction method of multipliers, ADMM)[27]求解得到每一个 $\boldsymbol{s}_j$. 使用 ADMM 算法的关键是写出该问题对应的增广拉格朗日函数, 如公式(7)所示:

$$L(\boldsymbol{s}_j,\boldsymbol{z}_j,\rho,\boldsymbol{Y})=\frac{1}{2}\|\boldsymbol{D}^{-j}\boldsymbol{s}_j-\boldsymbol{D}^j\|_2^2+\lambda\|\boldsymbol{z}_j\|_1+\frac{\rho}{2}\|\boldsymbol{s}_j-\boldsymbol{z}_j+\boldsymbol{Y}/\rho\|_2^2\tag{7}$$

其中, ρ是惩罚参数, 控制 $\boldsymbol{s}_j$ 与 $\boldsymbol{z}_j$ 的相近程度; $\boldsymbol{Y}$ 是拉格朗日乘子. 引入缩放对偶变量$\boldsymbol{\mu}=\boldsymbol{Y}/\rho$, 根据 ADMM 算法的更新规则, 求得 3 个变量 $\boldsymbol{s}_j$、$\boldsymbol{z}_j$、$\boldsymbol{\mu}$的更新公式如下:

$$\left.\begin{aligned}\boldsymbol{s}_j^{k+1}&=((\boldsymbol{D}^{-j})'\boldsymbol{D}^{-j}+\rho\boldsymbol{I})^{-1}((\boldsymbol{D}^{-j})'\boldsymbol{D}^j\boldsymbol{D}_j+\rho(\boldsymbol{z}_j^k-\boldsymbol{\mu}^k)),\\\boldsymbol{z}_j^{k+1}&=S_{\lambda/\rho}(\boldsymbol{s}_j^{k+1}+\boldsymbol{\mu}^k),\\\boldsymbol{\mu}^{k+1}&=\boldsymbol{\mu}^k+\boldsymbol{s}_j^{k+1}-\boldsymbol{z}_j^{k+1}\end{aligned}\right\}\tag{8}$$

其中, $(\cdot)'$为转置操作, S 为 $L1$ 范数的近邻算子(proximity operator), $S_\omega(t)=\max(t-\omega)+\min(t+\omega)$. 按照公式(8)更新3个变量至满足算法的迭代收敛条件, 得到最终的$\boldsymbol{s}_j$. 对每个$\boldsymbol{s}_j$, 在第j个位置插入0对其进行扩展, 得到$\boldsymbol{S}_j\in R^l$, 0 元素表示在标记相关性求解时不考虑标记自身产生的影响. 将得到的所有 $\boldsymbol{S}_j=[\boldsymbol{S}_1,\boldsymbol{S}_2,\ldots,\boldsymbol{S}_l]$按 j 的数值由小到大排列, 得到标记相关性矩阵 $\boldsymbol{S}=[\boldsymbol{S}_1,\boldsymbol{S}_2,\ldots,\boldsymbol{S}_l]\in R^{l\times l}$, 其中, $s_{ii}=0(i=1,2,\ldots,l)$.

2.2 标记分布学习模型训练

本节, 我们提出一种新的标记学习方法——考虑标记间协作的标记分布学习(LDLCL), 将第 2.1 节中学习到的标记相关性 $\boldsymbol{S}$ 加入到期望的预测模型中. 假设 $f(\boldsymbol{X})=[f_1(\boldsymbol{X}),f_2(\boldsymbol{X}),\ldots,f_l(\boldsymbol{X})]\in R^{n\times l}$为标记分布映射函数, $f_1(\boldsymbol{X})$, $f_2(\boldsymbol{X}),\ldots,f_l(\boldsymbol{X})$表示 l 个独立的标记分布预测算子, 同时考虑全部 l 个标签的预测结果, 根据公式(2), 我们得到如下公式(9):

$$(1-\alpha)f(\boldsymbol{X})+\alpha f(\boldsymbol{X})\boldsymbol{S}=f(\boldsymbol{X})((1-\alpha)I+\alpha\boldsymbol{S})\tag{9}$$

由公式(9)可以看出, 考虑标记间协作的标记分布学习问题可以看作两个独立的子问题: 训练原始的模型 $f(\boldsymbol{X})$, 利用标记相关性矩阵和模型的输出拟合最终预测结果.

为了将两个子问题结合起来, 我们引入一个标记独立的标记分布空间矩阵 $\boldsymbol{Z}\in R^{n\times l}$. 令 $\boldsymbol{G}=(1-\alpha)\boldsymbol{I}+\alpha\boldsymbol{S}$, $\boldsymbol{G}$ 表示标记整体的协作关系, 则考虑标记协作的预测结果为 $\boldsymbol{ZG}$, 并且预测的标记分布 $\boldsymbol{ZG}$ 要满足标记分布中描述度 $d_x^y\in[0,1]$且$\sum_y d_x^y=1$的约束, 由此可得如下目标函数:

$$\left.\begin{aligned}&\min_{\boldsymbol{Z},f}\frac{1}{2}\|f(\boldsymbol{X})-\boldsymbol{Z}\|_F^2+\frac{\lambda_1}{2}\|\boldsymbol{ZG}-\boldsymbol{D}\|_F^2+\frac{\lambda_2}{2}\Omega(f)\\&\text{s.t. }\boldsymbol{ZG}\times\boldsymbol{1}_{l\times1}=\boldsymbol{1}_{n\times1}\\&\boldsymbol{ZG}\geqslant\boldsymbol{0}_{n\times l}\end{aligned}\right\}\tag{10}$$

其中, 第 1 项为模型训练项, 使训练得到的模型尽可能地接近标记分布空间 $\boldsymbol{Z}$; 第 2 项为预测拟合项, 目的是使利用了标记间协作的预测结果更接近真实的标记分布; 第 3 项为模型参数控制项, 防止模型过拟合. λ_1、λ_2 为权重参数, 用于权衡模型训练和预测拟合的重要程度. 两个约束条件保证预测拟合得到的标记分布满足对每个样本所有描述度都非负, 且描述度之和为 1.

假设映射函数 $f(\boldsymbol{X})=\varphi(\boldsymbol{X})\boldsymbol{W}$, 其中, $\varphi(\cdot)$表示对输入数据进行空间变换, $\boldsymbol{W}$ 为特征映射的权重矩阵. 采用标记分布学习中常用的 $L2$ 范数来约束模型参数 $\boldsymbol{W}$, 即:

$$\Omega(f)=\|\boldsymbol{W}\|_F^2\tag{11}$$

将 $f(\boldsymbol{X})=\varphi(\boldsymbol{X})\boldsymbol{W}$ 和公式(11)代入公式(10), 得到最终的目标函数:

$$\left.\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{W}}\frac{1}{2}\|\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{Z}\|_F^2+\frac{\lambda_1}{2}\|\boldsymbol{ZG}-\boldsymbol{D}\|_F^2+\frac{\lambda_2}{2}\Omega(f)\\&\text{s.t. }\boldsymbol{ZG}\times\mathbf{1}_{l\times1}=\mathbf{1}_{n\times1}\\&\boldsymbol{ZG}\geqslant\mathbf{0}_{n\times l}\end{aligned}\right\}\tag{12}$$

2.3 优　化

接下来, 我们对第 2.2 节标记分布学习模型训练的目标函数公式(12)的求解进行具体说明. 公式(12)为含有两个未知变量 $\boldsymbol{Z}$ 和 $\boldsymbol{W}$ 的约束优化问题, 采用迭代更新法求解公式(12). 具体步骤如下.

1) 固定 $\boldsymbol{W}$ 更新 $\boldsymbol{Z}$

在 $\boldsymbol{W}$ 为定值时, 关于 $\boldsymbol{Z}$ 的函数为

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{Z}\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{W}\|_F^2\tag{13}$$

直接对公式(13)关于 $\boldsymbol{W}$ 求导令导数为 0, 可以求得最佳的未知变量 $\boldsymbol{W}^*$, 并且在之后的标记分布预测中使用 $f(\boldsymbol{X}_{test})=\varphi(\boldsymbol{X}_{test})\boldsymbol{W}^*$求得测试样本的标记分布. 但这样做需要显式地计算特征映射$\varphi(\boldsymbol{X})$. $\varphi(\cdot)$将特征从低维空间映射到高维空间, 直接计算映射结果计算量很大; 同时, 我们也不希望关注特征映射的具体形式. 因此, 为了进一步简化一般非线性情形下的核扩展, 我们引入一个新的变量 $\boldsymbol{E}=\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{Z}$, 使得可以使用核方法处理此问题, 把高维空间的计算通过低维空间的计算外加一些线性变换完成.

引入变量 $\boldsymbol{E}$ 后, 公式(12)转化为有 3 个未知变量的约束优化问题:

$$\left.\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{W},\boldsymbol{E}}\frac{1}{2}\|\boldsymbol{E}\|_F^2+\frac{\lambda_1}{2}\|\boldsymbol{ZG}-\boldsymbol{D}\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{W}\|_F^2\\&\text{s.t. }\boldsymbol{Z}-\varphi(\boldsymbol{X})\boldsymbol{W}=\boldsymbol{E}\\&\boldsymbol{ZG}\times\mathbf{1}_{l\times1}=\mathbf{1}_{n\times1}\\&\boldsymbol{ZG}\geqslant\mathbf{0}_{n\times l}\end{aligned}\right\}\tag{14}$$

固定 $\boldsymbol{Z}$ 更新 $\boldsymbol{W}$ 和 $\boldsymbol{E}$. 当 $\boldsymbol{Z}$ 为定值时, 关于 $\boldsymbol{W}$ 和 $\boldsymbol{E}$ 的拉格朗日函数为

$$L(\boldsymbol{E},\boldsymbol{W})=\|\boldsymbol{E}\|_F^2+\lambda_2\|\boldsymbol{W}\|_F^2+\langle\boldsymbol{A},\boldsymbol{Z}-\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{E}\rangle\tag{15}$$

对 $\boldsymbol{E}$、$\boldsymbol{A}$、$\boldsymbol{W}$ 分别求导, 得:

$$\left.\begin{aligned}\frac{\partial L}{\partial\boldsymbol{E}}=0\Rightarrow\boldsymbol{E}=\boldsymbol{A},\\\frac{\partial L}{\partial\boldsymbol{A}}=0\Rightarrow\boldsymbol{Z}-\varphi(\boldsymbol{X})\boldsymbol{W}=\boldsymbol{E},\\\frac{\partial L}{\partial\boldsymbol{W}}=0\Rightarrow\boldsymbol{W}=\frac{1}{\lambda_2}\varphi(\boldsymbol{X})'\boldsymbol{A}\end{aligned}\right\}\tag{16}$$

由公式(16)可得, $\boldsymbol{E}$ 和 $\boldsymbol{W}$ 两个变量都是关于 $\boldsymbol{A}$ 的函数. 定义 $\boldsymbol{H}=\frac{1}{\lambda_2}\boldsymbol{K}+\boldsymbol{I}$, 可以得到 $\boldsymbol{A}=\boldsymbol{H}^{-1}\boldsymbol{Z}$. 其中, $\boldsymbol{K}=\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'$为核函数, $\boldsymbol{I}$ 为单位阵. 本文中, 我们选高斯核函数计算 $\boldsymbol{K}$, 也就是说, $k_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2/(2\sigma^2))$. 核函数使我们无需显式计算每一个$\varphi(\boldsymbol{x}_i)$, 直接求出$\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_i)'$的值. 关于本文中所选高斯核函数中参数$\sigma$的取值, 将在第 4.3 节参数设置中详细说明.

2) 固定 $\boldsymbol{W}$ 更新 $\boldsymbol{Z}$

令模型输出 $f(\boldsymbol{X})=\varphi(\boldsymbol{X})\boldsymbol{W}=\frac{1}{\lambda_2}\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{A}=\frac{1}{\lambda_2}\boldsymbol{KA}=\boldsymbol{T}$, 用 $\boldsymbol{T}$ 替换公式(12)中的$\varphi(\boldsymbol{X})\boldsymbol{W}$, 当 $\boldsymbol{W}$ 为定值时, 关于 $\boldsymbol{Z}$ 的目标函数可写为公式(17):

$$\min_{\boldsymbol{Z},\boldsymbol{Y}_1,\boldsymbol{Y}_2}\|\boldsymbol{Z}-\boldsymbol{T}\|_2^2+\lambda_1\|\boldsymbol{ZG}-\boldsymbol{D}\|_F^2+\langle\boldsymbol{Y}_1,\boldsymbol{Z}-\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{E}\rangle+\langle\boldsymbol{Y}_2,\boldsymbol{ZG}\times\mathbf{1}_{l\times1}-\mathbf{1}_{n\times1}\rangle\tag{17}$$

使用对偶梯度法更新变量 $\boldsymbol{Z}$、$\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$:

$$\left.\begin{aligned}\boldsymbol{Z}&=(\boldsymbol{T}+\lambda_1\boldsymbol{D}\boldsymbol{G}'-\boldsymbol{Y}_1\boldsymbol{1}'_{l\times1}\boldsymbol{G}'-\boldsymbol{Y}_2\boldsymbol{G}')(\boldsymbol{I}+\lambda_1\boldsymbol{G}\boldsymbol{G}')^{-1},\\ \boldsymbol{Y}_1^{k+1}&=\boldsymbol{Y}_1^{k}+\alpha(\boldsymbol{Z}-\varphi(\boldsymbol{X})\boldsymbol{W}-\boldsymbol{E}),\\ \boldsymbol{Y}_2^{k+1}&=\boldsymbol{Y}_2^{k}+\alpha(\boldsymbol{Z}\boldsymbol{G}\times\boldsymbol{1}_{l\times1}-\boldsymbol{1}_{n\times1})\end{aligned}\right\} \tag{18}$$

其中, α为梯度上升的步长, $\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 为拉格朗日乘子.

根据公式(16)、公式(18)迭代更新变量 $\boldsymbol{W}$、$\boldsymbol{E}$、$\boldsymbol{Z}$ 值, 直到满足终止条件, 最后得到变量 $\boldsymbol{A}$ 的最优值.

3 标记分布预测

训练阶段得到最终的拉格朗日乘子 $\boldsymbol{A}$ 后, 根据 $f(\boldsymbol{X})=\varphi(\boldsymbol{X})\boldsymbol{W}=\frac{1}{\lambda_2}\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{A}=\frac{1}{\lambda_2}\boldsymbol{K}\boldsymbol{A}$ 计算得到模型输出,再根据 $f(\boldsymbol{X})((1-\alpha)\boldsymbol{I}+\alpha\boldsymbol{S})=f(\boldsymbol{X})\boldsymbol{G}$ 得到最终的预测结果, 即公式(19)所示:

$$f(\boldsymbol{X}_{test})=\frac{1}{\lambda_2}\boldsymbol{K}_{test}\boldsymbol{A}\boldsymbol{G} \tag{19}$$

我们使用 $\boldsymbol{X}_{train}$ 表示训练集样本, $\boldsymbol{X}_{test}$ 表示测试集样本. 需要注意的是: 在训练阶段, $\boldsymbol{K}=\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'$中的 $\boldsymbol{X}$ 为训练集样本 $\boldsymbol{X}_{train}$, 用以度量训练集样本间的相似度. 在标记预测过程中, $\boldsymbol{K}_{test}=\varphi(\boldsymbol{X}_{test})\varphi(\boldsymbol{X}_{train})'\in R^{m\times n}$, 用来度量测试集和训练集样本间的相似度. 其中, m 为测试集样本个数, n 为训练集样本个数. 通过使用核函数, 使我们可以不用关心特征映射空间$\varphi(\cdot)$的具体形式, 直接计算样本间的相似度.

图 3 给出了 LDLCL 算法训练和测试过程的整体流程图(样本相似度 [1] 为训练集样本间的相似度矩阵 $\boldsymbol{K}$, 样本相似度 [2] 为训练集和测试集样本间的相似度矩阵 $\boldsymbol{K}_{test}$).

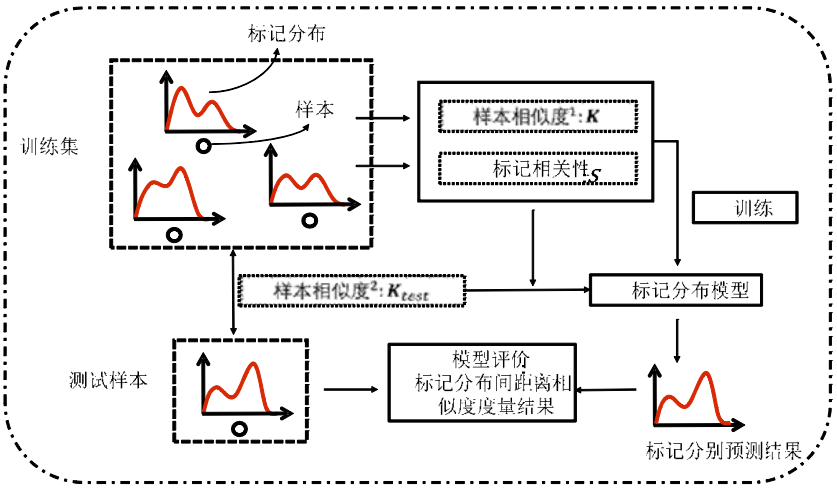


图 3 LDLCL 算法流程图

4 实 验

4.1 数据集

我们的 14 个数据集来自标记分布学习网站(http://LDL.herokuapp.com/download). Yeast_xx 系列 10 个数据集(Yeast-alpha, Yeast-cdc, Yeast-elu, Yeast-diau, Yeast-heat, Yeast-cold, Yeast-dtt, Yeast-spo, Yeast-spo5 和 Yeast-spoem)来自酿酒酵母上的生物学实验, 数据集一共包含 2 465 个酵母菌基因样本, 每个基因通过 24 个特征表示, 每个样本对应的标记分布是不同时间点上基因的表达水平. Natural Scene 数据集由 2 000 幅自然场景图像组成, 研究人员要求 10 个标注者用 9 种可能的标签给这些图像贴上标签, 即植物、天空、云、雪、建筑、沙漠、山、水和太阳. 对于每个图像, 每个人选择相关的标签并独立地按降序排列, 因此多标记排序的结果高度

不一致. 使用非线性规划过程将不一致的排序转换为标记分布[28], 最后利用文献[29]中提出的方法, 为每个图像提取 294 维特征向量作为图像的特征表示. 数据集 S-JAFFE 和 SBU_3DFE 是人类表情图像数据集, 分别包含 213 张面部表情图像和 2 500 张 3D 面部表情图像, 每张图像共有 243 个特征, 标记数 6 代表人类的 6 种表情, 包括快乐、悲伤、惊讶、恐惧、愤怒和厌恶. 多个不同的人根据图片中的人脸表情对 6 种表情进行打分, 最后对分数进行归一化得到标记分布. 数据集 Movie 是关于电影的用户评分, 一共包含 7 755 部电影, 每部电影共有 1 869 个特征, 每部电影的评分有 5 个级别, 相当于有 5 个标记. 将每个级别的评分人数占总评分人数的比例作为对应标记的描述度, 生成一个标记分布. 表 1 中总结了这 14 个数据集的一些简要统计信息.

表 1 实验数据集统计信息表

ID	Dataset	#Instance	#Feature	#Label
1	Yeast-alpha	2 465	24	18
2	Yeast-cdc	2 465	24	15
3	Yeast-elu	2 465	24	14
4	Yeast-diau	2 465	24	7
5	Yeast-heat	2 465	24	6
6	Yeast-cold	2 465	24	4
7	Yeast-dtt	2 465	24	4
8	Yeast-spo	2 465	24	6
9	Yeast-spo5	2 465	24	3
10	Yeast-spoem	2 465	24	2
11	Natural Scene	2 000	294	9
12	S-JAFFE	213	243	6
13	SBU_3DFE	2 500	243	6
14	Movie	7 755	1 869	5

4.2 评价指标

标记分布学习算法的输出是标记分布, 评价算法表现的一个自然指标就是预测的标记分布与真实分布之间的平均距离或相似度. 根据文献[6]中的建议, 本实验中采用 6 种代表性的标记分布评价指标对算法性能进行验证, 包含两个分布之间的 4 个距离度量: Chebyshev 距离(Chev↓)、Clark 距离(Clark↓)、Canberra 距离(Canberra↓)、Kullback-Leibler 散度(KL-div↓)和两个分布之间的两个相似性度量: 余弦相关系数(Cosine↑)和交叉相似度(Intersec↑). ↓表示评价指标度量值越低性能越好, ↑表示评价指标度量值越高性能越好.

4.3 实验设置

我们将提出的 LDLCL 算法与 7 种算法进行了比较: PT-SVM 算法、PT-Bayes 算法、AA-BP 算法、SA-IIS 算法、SA-BFGS 算法、LDLLC 算法[11]和 LDL-LCLR 算法[12]. 其中, 前 5 种算法为经典的 LDL 算法, 后两种算法(LDLLC 和 LDL-LCLR)是近期 LDL 领域考虑标记间相关性的算法. 所有算法的代码均来自于原作者共享, 实验中参数取值均按照相应文献中的建议.

下面对 LDLCL 中使用核函数计算样本间相似度时参数的取值给出详细说明.

本文中, 我们使用高斯核函数计算样本间相似度, 也就是说, 两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的相似度按照 $k_{ij}=K(\boldsymbol{x}_i, \boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2/(2\sigma^2))$ 计算. 在使用高斯核函数时, σ控制样本间的分离程度/差别. σ越小, 核函数对 x 的衰减越快, 意味着放大了数据 x 之间的差别, 即高斯核 $K(x)$对 x 值的变化很敏感. 大多数研究中, 对σ简单取值为所研究样本间欧式距离的平均值, 即公式(20)所示:

$$\sigma_1=\frac{n(n-1)}{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sqrt{\sum_{k=1}^{d}(\boldsymbol{x}_i^k-\boldsymbol{x}_j^k)^2} \tag{20}$$

其中, n 为样本个数, d 为样本特征维数, $\boldsymbol{x}_i^k$ 表示样本 $\boldsymbol{x}_i$ 对应的第 k 维特征.

对数据集 Natural Scene、S-JAFFE、SBU_3DFE 和 Movie, 参照大多数研究, 我们取$\sigma=\sigma_1$.

对 Yeast_xx 系列 10 个数据集, 考虑到它们是相同样本的不同标记分布表示, 在计算样本间相似度时, 我们加入了标记个数对样本相似度的影响. 具体来讲, 我们取$\sigma=\sigma_1\times\omega\times num_labels$, 其中, ω为权重参数, 用来控制标记个数对σ的影响程度; *num_labels* 对应不同数据集包含的标记个数, 本文中我们取ω=0.5. 对具有不同标

记个数的相同样本计算相似度时考虑标记个数是合理的, 样本的标记个数越少, 样本间相同标记的数目相应也会越少, 样本间的差别就会越明显, 样本间的相似度值对样本变化也就越敏感, 所需要的σ值就越小. 表 2 列出了在 10 个数据集上是否考虑标记个数的实验结果, 对每个数据集, 第 1 行为考虑标记个数的实验结果, 第 2 行为不考虑标记个数的实验结果, 我们用粗体标记每个度量的最佳结果. 由表 2 可知: 在计算具有不同标记个数的相同样本间相似性时, 考虑标记个数的影响是很有必要的.

表 2 考虑标记个数与否的实验结果

Dataset	KL-div	Chev	Canberra	Clark	Intersec	Cosine
Yeast-alpha	**0.005 464**	**0.013 408**	**0.679 857**	**0.209 443**	**0.962 472**	**0.994 626**
	0.005 763	0.013 691	0.702 273	0.215 437	0.961 215	0.994 332
Yeast-cdc	**0.006 943**	**0.016 176**	**0.645 681**	**0.215 351**	**0.957 519**	**0.993 332**
	0.007 344	0.016 539	0.668 669	0.221 883	0.955 989	0.992 937
Yeast-elu	**0.006 145**	**0.016 242**	**0.581 670**	**0.198 559**	**0.958 971**	**0.994 061**
	0.006 525	0.016 507	0.600 815	0.203 811	0.957 608	0.993 688
Yeast-diau	**0.012 841**	**0.036 577**	**0.425 697**	**0.198 323**	**0.940 992**	**0.988 184**
	0.013 500	0.037 029	0.434 806	0.201 783	0.939 674	0.987 535
Yeast-heat	**0.012 377**	**0.041 773**	**0.360 579**	**0.180 791**	**0.940 843**	**0.988 230**
	0.012 844	0.042 320	0.367 369	0.183 681	0.939 731	0.987 785
Yeast-cold	**0.012 055**	**0.050 726**	**0.238 671**	**0.138 604**	**0.941 205**	**0.988 672**
	0.012 371	0.051 312	0.241 847	0.140 267	0.940 398	0.988 359
Yeast-dtt	**0.006 172**	**0.035 696**	**0.167 429**	**0.097 264**	**0.958 696**	**0.994 171**
	0.006 360	0.036 184	0.169 655	0.098 543	0.958 134	0.993 986
Yeast-spo	**0.024 466**	**0.058 130**	**0.511 758**	**0.248 994**	**0.915 772**	**0.977 083**
	0.026 379	0.059 859	0.527 706	0.256 014	0.913 085	0.975 285
Yeast-spo5	**0.028 680**	**0.090 499**	**0.280 452**	**0.182 666**	**0.909 501**	**0.974 658**
	0.029 528	0.092 030	0.285 202	0.185 659	0.907 970	0.973 846
Yeast-spoem	0.025 349	0.088 715	0.183 366	0.131 715	0.911 285	0.978 308
	0.025 310	**0.088 669**	**0.183 280**	**0.131 656**	**0.911 331**	**0.978 343**

4.4 实验结果

对于每一个数据集, 本文中, 我们都采用了十折交叉验证(ten-fold cross validation)进行实验. 具体来说, 实验中, 每个数据集中的实例都被随机地分为 10 部分: 一部分用于测试, 其余部分用于训练, 共进行 10 次实验. 我们记录了每个算法在 6 个评价指标的结果, 实验结果以“平均值±标准差(等级)(mean±std. (rank))”的形式给出, 等级是指所有 LDL 算法对每个度量的预测效果的排序. 此外, 在每个表中, 我们用粗体标记每个度量的最佳结果. 由第 4.2 节可知, 6 个评价指标可分为两类: 4 个距离度量评价指标(Chev↓, Clark↓, Canberra↓, KL-div↓)和两个相似性度量评价指标(Cosine↑, Intersec↑), 相同类型的评价指标实验结果相似. 因此, 本文在每一类中选择一半评价指标共 3 个评价指标(Clark↓, Canberra↓, Intersec↑)列出实验结果. 实验结果见表 3–表 5, 每个表展示一个评估度量的比较结果.

表 3 不同 LDL 算法的 Intersec 测量结果(平均值±标准差(等级))比较

ID	PT-SVM	PT-Bayes	AA-BP	SA-IIS
1	0.9597±0.001(4)	0.7762±0.007(8)	0.8754±0.010(7)	0.9422±0.001(6)
2	0.9556±0.001(3)	0.7757±0.014(7)	0.8904±0.005(6)	0.9396±0.000(5)
3	0.9564±0.001(4)	0.7799±0.009(8)	0.8948±0.005(7)	0.9404±0.001(6)
4	0.9289±0.006(4)	0.7731±0.012(7)	0.9202±0.003(6)	0.9257±0.001(5)
5	0.9379±0.002(3)	0.7707±0.010(7)	0.9251±0.003(5)	0.9234±0.001(6)
6	0.9360±0.004(4)	0.7880±0.009(7)	0.9339±0.002(5)	0.9291±0.001(6)
7	0.9567±0.001(4)	0.8023±0.010(8)	0.9485±0.003(7)	0.9424±0.001(6)
8	0.9066±0.005(5)	0.7752±0.012(7)	0.9090±0.005(4)	0.9061±0.002(6)
9	0.9073±0.004(4)	0.7974±0.010(8)	0.9070±0.002(5)	0.9028±0.003(7)
10	0.9075±0.006(6)	0.8171±0.020(8)	0.9030±0.006(7)	0.9075±0.004(5)
11	0.3453±0.051(8)	0.3476±0.007(7)	0.4955±0.014(4)	0.4619±0.010(6)
12	0.8427±0.010(7)	0.8460±0.060(6)	0.8280±0.021(8)	0.8546±0.015(4)
13	0.8343±0.006(8)	0.8383±0.004(6)	0.8348±0.006(7)	0.8390±0.005(5)
14	0.6916±0.038(8)	0.7230±0.003(7)	0.7990±0.007(5)	0.8043±0.002(4)
Avg. Rank	5.14	7.21	5.93	5.50

表 3 不同 LDL 算法的 Intersec 测量结果(平均值±标准差(等级))比较(续)

ID	SA-BFGS	LDLLC	LDL-LCLR	Ours
1	0.9501±0.002(5)	0.9623±0.000(3)	0.9624±0.000(2)	**0.9625±0.000(1)**
2	0.9498±0.001(4)	0.9574±0.000(2)	0.9574±0.000(2)	**0.9575±0.000(1)**
3	0.9488±0.002(5)	0.9588±0.000(3)	0.9589±0.000(2)	**0.9590±0.000(1)**
4	0.9337±0.000(3)	0.9403±0.000(2)	0.9403±0.000(2)	**0.9410±0.000(1)**
5	0.9340±0.001(4)	0.9402±0.000(2)	0.9402±0.000(2)	**0.9408±0.000(1)**
6	0.9372±0.001(3)	0.9408±0.000(2)	0.9408±0.000(2)	**0.9412±0.000(1)**
7	0.9540±0.001(5)	0.9583±0.000(3)	0.9583±0.000(2)	**0.9587±0.000(1)**
8	0.9104±0.000(3)	0.9155±0.000(2)	0.9155±0.000(2)	**0.9158±0.000(1)**
9	0.9029±0.001(6)	0.9087±0.000(2)	0.9086±0.000(3)	**0.9095±0.000(1)**
10	0.9114±0.000(3)	0.9132±0.000(2)	**0.9131±0.000(1)**	0.9113±0.000(4)
11	0.4623±0.008(5)	0.5508±0.000(3)	0.5557±0.001(2)	**0.6004±0.001(1)**
12	0.8613±0.006(3)	0.8487±0.000(5)	0.8719±0.000(2)	**0.8900±0.001(1)**
13	0.8454±0.008(3)	0.8412±0.000(4)	0.8515±0.000(2)	**0.8627±0.000(1)**
14	0.7701±0.004(6)	0.8324±0.000(2)	0.8223±0.001(3)	**0.8383±0.000(1)**
Avg. Rank	4.14	2.64	2.07	**1.21**

表 4 不同 LDL 算法的 Clark 测量结果(平均值±标准差(等级))比较

ID	PT-SVM	PT-Bayes	AA-BP	SA-IIS
1	0.2238±0.004(3)	1.1541±0.034(7)	0.7236±0.020(6)	0.3053±0.006(5)
2	0.2234±0.007(3)	1.0601±0.066(7)	0.5728±0.030(6)	0.2932±0.004(5)
3	0.2097±0.005(4)	1.0050±0.041(8)	0.5246±0.028(7)	0.2751±0.006(6)
4	0.2376±0.018(5)	0.7487±0.042(8)	0.2677±0.010(7)	0.2409±0.006(6)
5	0.1893±0.006(5)	0.6829±0.026(7)	0.2261±0.010(6)	0.2260±0.005(5)
6	0.1496±0.010(5)	0.5149±0.024(8)	0.1552±0.005(6)	0.1643±0.004(7)
7	0.1023±0.003(4)	0.4807±0.040(8)	0.1206±0.008(6)	0.1332±0.003(7)
8	0.2736±0.011(6)	0.6686±0.040(7)	0.2950±0.010(6)	0.2759±0.006(5)
9	0.1867±0.009(5)	0.4220±0.020(7)	0.1870±0.005(4)	0.1944±0.009(5)
10	0.1366±0.008(5)	0.3065±0.030(8)	0.1890±0.012(7)	0.1367±0.007(6)
11	2.5705±0.029(8)	2.5259±0.015(7)	2.4534±0.018(4)	2.4703±0.019(5)
12	0.4419±0.025(7)	0.4327±0.021(6)	0.5164±0.072(8)	0.4082±0.037(3)
13	0.4263±0.013(7)	0.4137±0.010(4)	0.4454±0.020(8)	0.4156±0.012(6)
14	0.8343±0.068(8)	0.8044±0.010(7)	0.6533±0.010(6)	0.5783±0.007(5)
Avg. Rank	5.36	7.07	6.21	5.43
ID	SA-BFGS	LDLLC	LDL-LCLR	Ours
1	0.2689±0.008(4)	0.2102±0.000(2)	0.2102±0.001(2)	**0.2094±0.000(1)**
2	0.2477±0.007(4)	0.2158±0.000(2)	0.2158±0.000(2)	**0.2154±0.000(1)**
3	0.2438±0.008(5)	0.1992±0.000(3)	0.1990±0.000(2)	**0.1986±0.000(1)**
4	0.2201±0.002(4)	0.2006±0.000(2)	0.2008±0.000(3)	**0.1983±0.000(1)**
5	0.1998±0.003(4)	0.1826±0.000(2)	0.1826±0.000(2)	**0.1808±0.000(1)**
6	0.1471±0.004(4)	0.1396±0.000(3)	0.1395±0.000(2)	**0.1386±0.000(1)**
7	0.1084±0.003(5)	0.0983±0.000(3)	0.0982±0.000(2)	**0.0973±0.000(1)**
8	0.2639±0.003(3)	0.2498±0.000(2)	0.2498±0.000(2)	**0.2490±0.000(1)**
9	0.1962±0.001(6)	0.1841±0.000(2)	0.1841±0.000(2)	**0.1827±0.000(1)**
10	0.1312±0.001(3)	**0.1292±0.000(1)**	0.1293±0.000(2)	0.1317±0.000(4)
11	2.4754±0.013(6)	2.4456±0.000(2)	**2.4301±0.001(1)**	2.4459±0.001(3)
12	0.4103±0.019(4)	0.4255±0.001(5)	0.3647±0.001(2)	**0.3273±0.004(1)**
13	0.3984±0.016(3)	0.4091±0.000(5)	0.3863±0.000(2)	**0.3603±0.000(1)**
14	0.5750±0.011(4)	**0.5227±0.001(1)**	0.5516±0.002(3)	0.5355±0.007(2)
Avg. Rank	4.21	2.50	2.07	**1.43**

表 5 不同 LDL 算法的 Canberra 测量结果(平均值±标准差(等级))比较

ID	PT-SVM	PT-Bayes	AA-BP	SA-IIS
1	0.7299±0.015(4)	4.1371±0.138(8)	2.3906±0.218(7)	1.0233±0.020(6)
2	0.6371±0.023(4)	3.4657±0.235(8)	1.7352±0.090(7)	0.8997±0.013(6)
3	0.6184±0.015(4)	3.1789±0.148(8)	1.5429±0.070(7)	0.8254±0.019(6)
4	0.5103±0.041(5)	1.6832±0.093(8)	0.5756±0.025(7)	0.5267±0.014(6)
5	0.3778±0.012(5)	1.4394±0.062(7)	0.4560±0.020(5)	0.4599±0.011(6)
6	0.2584±0.018(6)	0.9062±0.045(8)	0.2679±0.010(7)	0.2541±0.000(5)
7	0.1757±0.005(4)	0.8443±0.080(8)	0.2081±0.010(6)	0.2308±0.006(7)
8	0.5632±0.027(5)	1.4094±0.084(8)	0.6040±0.030(7)	0.5658±0.011(6)
9	0.2867±0.014(3)	0.6555±0.040(7)	0.2878±0.008(4)	0.2994±0.013(5)

表 5 不同 LDL 算法的 Canberra 测量结果(平均值±标准差(等级))比较(续)

ID	PT-SVM	PT-Bayes	AA-BP	SA-IIS
10	0.1905±0.012(6)	0.4117±0.040(8)	0.2910±0.018(7)	0.1905±0.009(5)
11	7.2839±0.152(8)	7.1627±0.063(7)	6.7498±0.078(4)	6.8103±0.085(5)
12	0.9231±0.054(7)	0.9054±0.054(6)	1.0510±0.143(8)	0.8503±0.081(4)
13	0.9210±0.032(7)	0.9020±0.023(6)	0.9367±0.039(8)	0.8995±0.030(5)
14	1.6053±0.167(8)	1.5590±0.020(7)	1.2507±0.020(5)	1.1119±0.014(4)
Avg. Rank	5.43	7.43	6.36	5.43
ID	SA-BFGS	LDLLC	LDL-LCLR	Ours
1	0.9012±0.032(5)	0.6829±0.000(3)	0.6819±0.000(2)	**0.6799±0.000(1)**
2	0.7582±0.026(5)	0.6469±0.000(2)	0.6474±0.000(3)	**0.6457±0.000(1)**
3	0.7242±0.027(5)	0.5838±0.000(3)	0.5829±0.000(2)	**0.5817±0.000(1)**
4	0.4749±0.006(4)	0.4307±0.000(2)	0.4310±0.000(3)	**0.4257±0.000(1)**
5	0.4018±0.008(4)	0.3641±0.000(2)	0.3641±0.000(2)	**0.3606±0.000(1)**
6	0.2539±0.006(4)	0.2403±0.000(3)	0.2402±0.000(2)	**0.2387±0.000(1)**
7	0.1863±0.005(5)	0.1691±0.000(3)	0.1690±0.000(2)	**0.1674±0.000(1)**
8	0.5429±0.005(4)	0.5132±0.000(2)	0.5134±0.000(3)	**0.5118±0.000(1)**
9	0.3012±0.003(6)	0.2828±0.000(2)	0.2828±0.000(2)	**0.2805±0.000(1)**
10	0.1828±0.001(3)	**0.1798±0.000(1)**	0.1799±0.000(2)	0.1834±0.000(4)
11	6.8247±0.058(6)	6.6942±0.001(3)	6.6606±0.000(2)	**6.6211±0.005(1)**
12	0.8487±0.033(3)	0.8880±0.001(5)	0.7544±0.002(2)	**0.6612±0.007(1)**
13	0.8596±0.046(3)	0.8859±0.000(4)	0.8290±0.000(2)	**0.7640±0.000(1)**
14	1.2944±0.023(6)	**1.0049±0.002(1)**	1.0629±0.004(3)	1.0086±0.001(2)
Avg. Rank	4.50	2.57	2.29	**1.29**

4.5 对比实验

为了验证 LDLCL 算法中考虑标记相关性的必要性，我们不考虑标记间的相关性，即令 $\boldsymbol{S}=\boldsymbol{I}$，其他实验参数设置保持不变，进行了对比实验. 若在标记分布学习模型训练过程中不考虑标记间相关性，公式(9)可写为

$$(1-\alpha)f(\boldsymbol{X})+\alpha f(\boldsymbol{X})\boldsymbol{S}=f(\boldsymbol{X}) \tag{21}$$

此时，训练阶段的目标函数应去掉参数 $\boldsymbol{Z}$ 改写为

$$\left.\begin{aligned}&\min_{\boldsymbol{E},\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{E}\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{W}\|_F^2\\&\text{s.t. } D-\varphi(\boldsymbol{X})\boldsymbol{W}=\boldsymbol{E}\end{aligned}\right\} \tag{22}$$

与第 2.3 节优化中类似，关于 $\boldsymbol{E}$、$\boldsymbol{W}$ 的拉格朗日函数为 $L(\boldsymbol{E},\boldsymbol{W})=\frac{1}{2}\|\boldsymbol{E}\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{W}\|_F^2+\langle \boldsymbol{Y},\boldsymbol{D}-\varphi(\boldsymbol{X})W-\boldsymbol{E}\rangle$.

对变量 $\boldsymbol{E}$、$\boldsymbol{W}$、$\boldsymbol{A}$ 分别求导，可以得到 $\boldsymbol{A}=\boldsymbol{H}^{-1}\boldsymbol{D}$. 得到 $\boldsymbol{A}$ 后，标记预测方式和第 3 节相同，其中，$\boldsymbol{G}=(1-\alpha)\boldsymbol{I}+\alpha\boldsymbol{S}=\boldsymbol{I}$. 表 6 为是否考虑标记相关性的实验结果，对每个数据集，第 1 行为考虑标记相关性的实验结果，第 2 行为不考虑标记相关性的实验结果，我们用粗体标记每个度量的最佳结果. 由表 6 可知：对大多数数据集，考虑标记相关性可以提高算法性能.

表 6 考虑标记相关性与否的实验结果

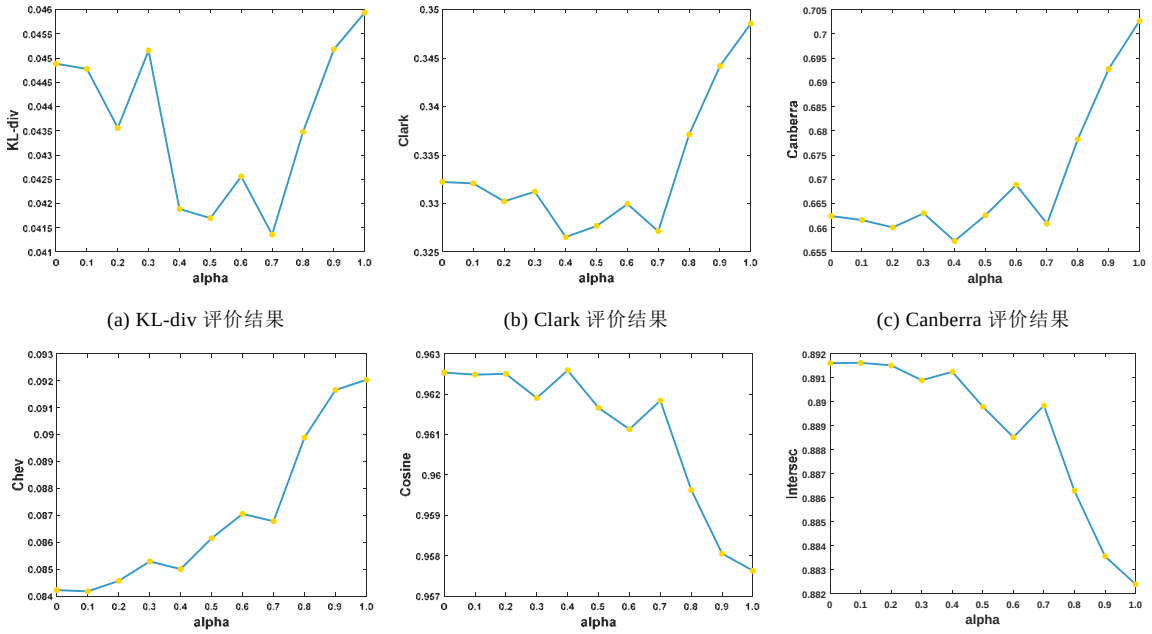
Dataset	KL-div	Chev	Canberra	Clark	Intersec	Cosine
Yeast-alpha	**0.005 464**	**0.013 408**	**0.679 857**	**0.209 443**	**0.962 472**	**0.994 626**
	0.005 477	0.013 421	0.680 892	0.209 709	0.962 414	0.994 614
Yeast-cdc	**0.006 943**	**0.016 176**	**0.645 681**	**0.215 351**	**0.957 519**	**0.993 332**
	0.006 954	0.016 187	0.646 107	0.215 501	0.957 491	0.993 321
Yeast-elu	**0.006 145**	**0.016 242**	**0.581 670**	**0.198 559**	**0.958 971**	**0.994 061**
	0.006 148	0.016 240	0.581 744	0.198 575	0.958 966	0.994 058
Yeast-diau	0.012 841	0.036 577	0.425 697	0.198 323	0.940 992	0.988 184
	0.012 806	**0.036 505**	**0.424 856**	**0.197 902**	**0.941 099**	**0.988 212**
Yeast-heat	0.012 377	0.041 773	0.360 579	0.180 791	0.940 843	0.988 230
	0.012 337	**0.041 649**	**0.360 257**	**0.180 485**	**0.940 902**	**0.988 268**
Yeast-cold	**0.012 055**	**0.050 726**	**0.238 671**	**0.138 604**	**0.941 205**	**0.988 672**
	0.012 292	0.051 250	0.241 240	0.139 957	0.940 563	0.988 442
Yeast-dtt	**0.006 172**	**0.035 696**	**0.167 429**	**0.097 264**	**0.958 696**	**0.994 171**
	0.006 288	0.036 077	0.169 262	0.098 285	0.958 236	0.994 055

表 6 考虑标记相关性与否的实验结果(续)

Dataset	KL-div	Chev	Canberra	Clark	Intersec	Cosine
Yeast-spo	**0.024 466**	**0.058 130**	**0.511 758**	**0.248 994**	**0.915 772**	**0.977 083**
	0.024 507	0.058 156	0.512 168	0.249 149	0.915 706	0.977 041
Yeast-spo5	**0.028 680**	**0.090 499**	**0.280 452**	**0.182 666**	**0.909 501**	**0.974 658**
	0.029 019	0.091 284	0.282 841	0.184 261	0.908 716	0.974 325
Yeast-spoem	**0.025 349**	**0.088 715**	**0.183 366**	**0.131 715**	**0.911 285**	**0.978 308**
	0.026 236	0.089 734	0.185 844	0.133 601	0.910 266	0.977 321
Natural Scene	**0.850 666**	**0.281 185**	**6.621 109**	**2.445 852**	**0.600 355**	**0.778 940**
	0.999 496	0.283 467	6.668 615	2.458 813	0.598 816	0.770 231
S-JAFFE	**0.041 787**	**0.086 158**	**0.661 166**	**0.327 299**	**0.890 001**	**0.961 723**
	0.049 803	0.085 192	0.683 662	0.346 009	0.889 471	0.961 242
SBU_3DFE	0.061 160	0.115 089	0.763 984	0.360 262	0.862 664	0.939 487
	0.057 821	**0.109 260**	**0.741 923**	**0.352 547**	**0.867 360**	**0.943 942**
Movie	**0.111 477**	**0.112 263**	**1.008 573**	**0.535 492**	**0.838 535**	**0.936 541**
	0.132 085	0.117 279	1.048 186	0.559 736	0.832 914	0.932 274

4.6 参数影响

为了研究折衷参数α对实验结果的影响，我们将α在[0,1]范围内以 0.1 步长取值，即α=[0,0.1,0.2,0.3,…,1]. 在不同数据集上运行 LDLCL 算法，得到α不同取值下相应的 6 个评价指标的结果. 图 4 所示为在 S-JAFFE 数据集上α不同取值下 6 个评价指标的结果. 由图 4 可以看出，$\alpha\in$[0.4,0.6]时，算法性能较稳定.

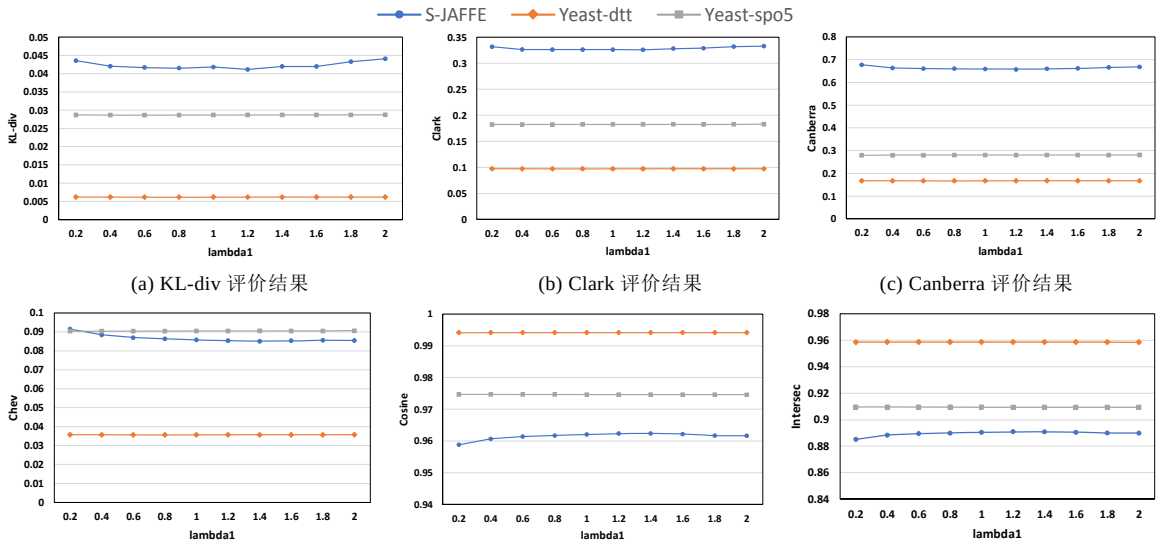


(a) KL-div 评价结果 (b) Clark 评价结果 (c) Canberra 评价结果

(d) Chev 评价结果 (e) Cosine 评价结果 (f) Intersec 评价结果

图 4 参数α在 LDLCL 中对数据集 S-JAFFE 的影响

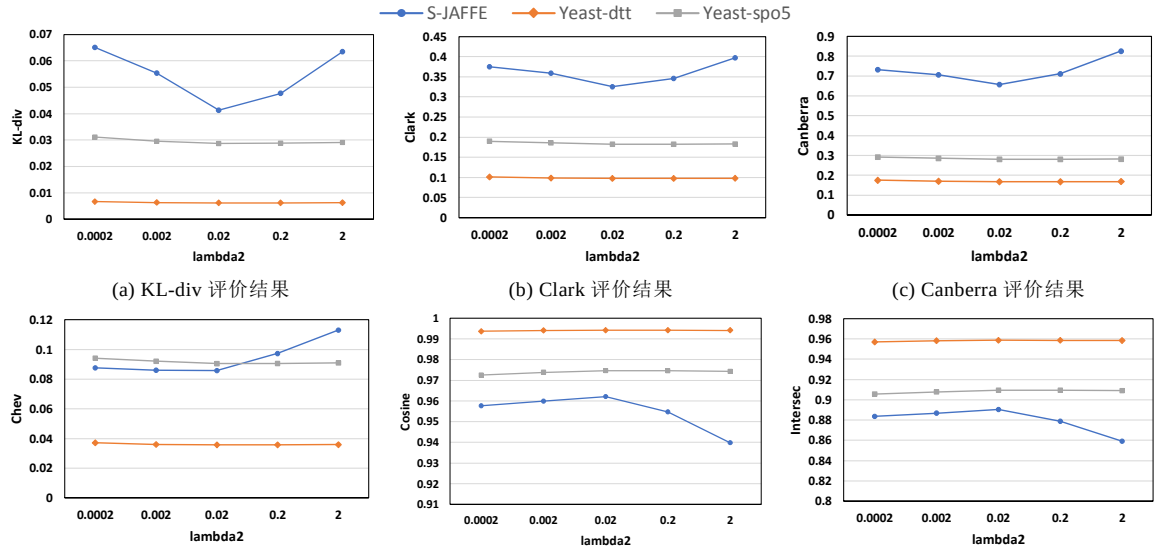
为了检验算法的鲁棒性，我们分析了权衡参数λ_1和λ_2对算法性能的影响. 我们在[0.2,2]范围内以 0.2 步长对λ_1进行选择，即λ_1=[0.2,0.4,0.6,0.8,1.0,1.2,1.4,1.6,1.8,2.0]；在[0.0002,2]范围内以 10 倍间隔对λ_2进行选择，即λ_1=[0.2,0.4,0.6,0.8,1.0,1.2,1.4,1.6,1.8,2.0]，λ_2=[0.0002,0.002,0.02,0.2,2]. 对所有的λ_1、λ_2，在不同数据集上运行 LDLCL 算法，得出相应取值下的结果. 图 5、图 6 所示分别为在 S-JAFFE、Yeast-dtt 和 Yeast-spo5 这 3 个数据集上，λ_1、λ_2不同取值下 6 个评价指标的实验结果.



(a) KL-div 评价结果 (b) Clark 评价结果 (c) Canberra 评价结果

(d) Chev 评价结果 (e) Cosine 评价结果 (f) Intersec 评价结果

图 5 参数λ_1在 LDLCL 中对数据集 S-JAFFE、Yeast-dtt、Yeast-spo5 的影响



(a) KL-div 评价结果 (b) Clark 评价结果 (c) Canberra 评价结果

(d) Chev 评价结果 (e) Cosine 评价结果 (f) Intersec 评价结果

图 6 参数λ_2在 LDLCL 中对数据集 S-JAFFE、Yeast-dtt、east-spo5 的影响

在图 5 中，令λ_2=0.02，只改变λ_1的取值．由图 5 可以看出：当λ_2固定、λ_1在[0.2,2]范围内改变时，算法在 6 个评价指标上的结果只在很小的范围内改变，说明算法对λ_1的取值鲁棒．在图 6 中，令λ_1=1，只改变λ_2的取值．由图 6 可以看出：当λ_1固定，对 S-JAFFE 数据集，λ_2=0.02 时，算法在 6 个评价指标上的实验结果最好；对 Yeast-dtt 和 Yeast-spo5 这两个数据集，λ_2的改变对实验结果影响很小，在λ_2=0.02 时，可以取得很好的实验结果．

实验中，对所有数据集，我们选择λ_1=1，λ_2=0.02，α=0.5.

5 总结和展望

本文中，我们将对于每个标记，最终的预测涉及到它自己的预测和其他标记的预测之间的协作这一关键

假设用于标记分布学习中. 基于这一假设, 我们通过稀疏重构得到了标记相关性矩阵, 并进一步将标记相关性用于标记分布学习模型的训练中, 在训练和预测中都显式利用了标记间的协作关系. 最后, 通过实验验证了我们这一方法的有效性. 如前所述, 标记间的相关性包含局部相关性和全局相关性, 本文中, 我们只考虑了标记间的全局相关性, 如何进一步利用标记间的局部相关性提高标记分布学习算法的性能, 有待后续研究.

References:

[1] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Trans. on Knowledge and Data Engineering, 2013, 26(8): 1819−1837.

[2] Gibaja E, Ventura S. Multi-label learning: A review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014, 4(6): 411−444.

[3] Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets. In: Gunopulos D, Turini F, Zaniolo C, Ramakrishnan N, Wu XD, eds. Proc. of the 8th IEEE Int'l Conf. on Data Mining. Pisa: IEEE Computer Society, 2008. 995−1000.

[4] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Machine Learning, 2011, 85(3): 333.

[5] Li Z, Tang J. Weakly supervised deep matrix factorization for social image understanding. IEEE Trans. on Image Processing, 2016, 6(1): 276−288.

[6] Geng X. Label distribution learning. IEEE Trans. on Knowledge and Data Engineering, 2016, 28(7): 1734−1748.

[7] Wang K, Geng X. Binary coding based label distribution learning. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI, 2018. 2783−2789.

[8] Zhou D, Zhang X, Zhou Y, Zhao Q, Geng X. Emotion distribution learning from texts. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 638−647.

[9] Plutchik R. A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H, eds. Proc. of the Theories of Emotion. Cambridge: Academic Press, 1980. 3−33

[10] Zhou Y, Xue H, Geng X. Emotion distribution recognition from facial expressions. In: Proc. of the 2015 ACM Multimedia Conf. Brisbane: ACM, 2015. 1247−1250.

[11] Jia X, Li W, Liu J, Zhang Y. Label distribution learning by exploiting label correlations. In: Liu CL, Zhang CS, Wang L, eds. Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018.

[12] Ren T, Jia X, Li W, Zhao S. Label distribution learning with label correlations via low-rank approximation. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI, 2019. 3325−3331.

[13] Zhang HR, Huang YT, Xu YY, Min F. COS-LDL: Label distribution learning by cosine-based distance-mapping correlation. IEEE Access, 2020, 8: 63961−63970.

[14] Zhao P, Zhou ZH. Label distribution learning by optimal transport. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018.

[15] Geng X, Yin C, Zhou ZH. Facial age estimation by learning from label distributions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2401−2412.

[16] Yin C, Geng X. Facial age estimation by conditional probability neural network. In: Liu CL, Zhang CS, Wang L, eds. Proc. of the Communications in Computer and Information Science. Beijing: Springer-Verlag, 2012. 243−250.

[17] Yang X, Gao BB, Xing C, Huo ZW, Wei XS, Zhou Y, Wu J, Geng X. Deep label distribution learning for apparent age estimation. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision Workshop. Santiago: IEEE Computer Society, 2015. 102−108.

[18] Gao BB, Xing C, Xie CW, Wu J, Geng X. Deep label distribution learning with label ambiguity. IEEE Trans. on Image Processing, 2017, 26(6): 2825−2838.

[19] Wang K, Geng X. Discrete binary coding based label distribution learning. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI, 2019. 3733−3739.

[20] Wang YB, Tian WQ, Cheng YS, Pei GS. Label distribution learning based on kernel extreme learning machine. Computer Engineering and Applications, 2018, 54(24): 128−135 (in Chinese with English abstract).

[21] Ren T, Jia X, Li W, Chen L, Li Z. Label distribution learning with label-specific features. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI, 2019. 3318−3324.

[22] Xu S, Shang L, Shen F. Latent semantics encoding for label distribution learning. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI, 2019. 3982−3988.

[23] Zhao Q, Geng X. Selection of target function in label distribution learning. Journal of Frontiers of Computer Science and Technology, 2017, 11(5): 708−719 (in Chinese with English abstract).

[24] Xing C, Geng X, Xue H. Logistic boosting regression for label distribution learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016. 4489−4497.

[25] Jia X, Li Z, Zheng X, Li W, Huang SJ. Label distribution learning with label correlations on local samples. IEEE Trans. on Knowledge and Data Engineering, 2019, 99: 1.

[26] Feng L, An B, He S. Collaboration based multi-label learning. In: Proc. of the 23rd AAAI Conf. on Artificial Intelligence, Vol. 33. Honolulu: AAAI, 2019. 3550−3557.

[27] Wei E, Ozdaglar A. Distributed alternating direction method of multipliers. In: Proc. of the 51st IEEE Conf. on Decision and Control (CDC). Maui: IEEE Computer Society, 2012. 5445−5450.

[28] Geng X, Luo L. Multilabel ranking with inconsistent rankers. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE Computer Society, 2014. 3742−3747.

[29] Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757−1771.

附中文参考文献:

[20] 王一宾, 田文泉, 程玉胜, 裴根生. 基于核极限学习机的标记分布学习. 计算机工程与应用, 2018, 54(24): 128−135.

[23] 赵权, 耿新. 标记分布学习中目标函数的选择. 计算机科学与探索, 2017, 11(5): 708−719.

李睿钰(1997－), 女, 学士, 主要研究领域为机器学习.

刘新媛(1996－), 女, 学士, 主要研究领域为机器学习.

祝继华(1982－), 男, 博士, 副教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习.