样本稀疏表达的标记分布学习算法

邵佳鑫，原盛，刘新媛，刘睿馨
（西安交通大学软件学院，710049，西安）

摘要：针对传统标记分布学习算法借助标记的全局相关性信息，忽略仅存于部分样本范围内标记局部相关性的问题，提出了一种基于样本稀疏表达的标记分布学习算法。借助样本点的自表达性质，建立稀疏表达优化模型，挖掘样本局部相关性信息。通过设计的标记分布目标函数约束，将得到的稀疏系数引入标记空间中，并将其作为隐含的标记空间局部相关性预测值，帮助标记分布模型的训练。使用交替方向乘子法求解样本稀疏系数，使用有限内存拟牛顿法求解标记分布目标函数，通过最大熵模型生成实例的标记分布预测值。在 11 个真实数据集上进行实验，并与 7 个现有标记分布学习算法进行对比。结果表明：所提算法在不同评价指标下的 55 次对比实验中取得了 1.52 的平均排名；面部表情数据集 SBU-3DFE 上，以相对熵衡量的表情判别准确度较标记分布学习问题转换算法 PT-SVM、适应性算法 AA-*k*NN 及专用算法 LDLLC 的分别提高了 3.10%、2.53%、2.48%；与传统标记分布学习算法相比，所提算法能够有效挖掘并利用标记局部相关性，具有良好的标记分布预测精度，且在不同类型的真实数据集上均能表现稳定。

关键词：标记分布学习；稀疏表达；最大熵模型
中图分类号：TP181 **文献标志码**：A
DOI：10.7652/xjtuxb202011017 **文章编号**：0253-987X(2020)11-0139-10

OSID 码

Label Distribution Learning via Sample Sparse Representation

SHAO Jiaxin, YUAN Sheng, LIU Xinyuan, LIU Ruixin
(School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: The traditional label distribution learning algorithms utilize the label correlation in a global way but ignore the local correlation. An algorithm called label distribution learning via sample sparse representation (LDL-SR) is proposed. Particularly, to capture the sample local correlation, LDL-SR uses the self-expressiveness property of samples to establish a sparse representation optimization model. Then, by the well-designed objective function, the obtained sparse coefficient is introduced into the label space to predict the correlation of the labels to help the training of the label distribution model. The sparse representation optimization model can be solved with alternating direction multiplier method (ADMM), and limited memory quasi-Newton method (L-BFGS) is chosen to optimize the target function of the label distribution model. Finally, LDL-SR calculates the predicted label distribution via the maximum entropy model. Compared with 7 existing label distribution learning algorithms, the experimental results on 11 real-world datasets show that this algorithm achieves an average ranking of 1.52 in 55 comparative experiments under different evaluation measures. Compared with the PT-SVM,

收稿日期：2020-05-07。 作者简介：邵佳鑫(1995—)，男，硕士生；原盛(通信作者)，男，讲师。 基金项目：国家自然科学基金资助项目(61902310)。
网络出版时间：2020-07-17 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20200715.1727.002.html

AA-kNN, and the LDLLC algorithms on facial expression dataset SBU-3DFE, the accuracy of expression discrimination is enhanced by 3.10%, 2.53% and 2.48% respectively. It is found that LDL-SR can achieve good prediction accuracy and stable performance on different real-world datasets.

Keywords: label distribution learning; sparse representation; maximum entropy model

在机器学习领域中,标记多义性学习是非常的火热的话题。目前,存在两种较为常见的机器学习范式来解决这个问题:单标记学习和多标记学习[1]。与单标记中一个实例只能关联一个标记不同,多标记学习能够给一个实例同时关联多个标记,为实例赋予更多的含义,因此多标记学习能够更加高效地解决实际问题。但是,多标记学习范式只能解决"标记能否描述该实例"的问题,无法解决"标记能多大程度描述该实例"这种更复杂的问题。为了解决这个问题,Geng 等提出一种标记分布学习(LDL)的新范式[2]。与多标记学习输出一个标记集合不同,标记分布学习输出的是一个标记分布,该分布中的每个分量为一个 0~1 之间的数字,表示该分量对应的标记与实例之间的相关程度。

LDL 提出之后,一些学者对其进行研究并提出了许多高效的算法。Geng 等为解决年龄预测问题,将单标记集转换为标记分布进行训练,提出 IIS-LDL 算法[2]。文献[3]基于神经网络提出条件概率神经网络算法。之后,Geng 等将 IIS-LLD 进行优化,结合拟牛顿优化算法,提出 BFGS-LDL 算法[4]。Wang 等通过深入分析标记分布模型的学习能力,得到标记分布模型和分类器模型的关系[5]。文献[6]在其理论基础上,通过绝对损失代替 KL 散度,采用信息熵为样本加权,提出用于分类的标记分布学习算法。

在特征选择方面:Xu 等为减轻噪声冗余特征的影响,利用潜在语义训练标记分布模型,并同时进行特征选择,提出 LSE-LDL 算法[7],Ren 等发现标记分布中一些标记的分量仅由其对应的部分特征决定,通过挖掘标记对应的特定特征帮助模型训练,提出 LDLSF 算法[8]。

此外,一些学者发现标记分布中标记间的相关性能够对模型训练起到积极的作用,即在一定条件下,标记分布中的某几个标记可能存在联系。例如,当一个面部表情实例在恐惧标记上有较高的描述度时,那么它更有可能在悲伤标记上同样有较高的描述度,而不是在高兴标记上。基于此:Zhou Y 等通过两标记间的皮尔森相关系数挖掘标记相关性,进行面部表情识别[9];Zhou D 等基于普鲁契克理论,利用标记相关性进行文本情感的标记分布学习[10];Jia 等引入距离映射函数表征标记相关性[11]。这些研究将标记间的相关性视为在整个样本空间共享的全局相关性,即所有实例对应标记分布中的标记拥有相同的相关性信息。但是,在实际问题中,标记间的相关性信息一般都是局部的,即一组实例共享相同标记相关性,不同组间的标记相关性存在差异,本文将这种相关性称为标记局部相关性。例如,对于自然风景图,理想的标记分布模型应当将风景图中的山、树等标记分布都正确识别。然而,草木茂盛的青山风景图中,山和树存在相关性,寸草不生的石山风景图中,山和树不存在相关性,即山标记和树标记具有的是局部相关性,而非全局相关性,若引入标记全局相关性就会误导模型判断。造成这种不同的原因是两个风景图实例并不属于同一个组,所以其标记间的相关性不同。

为了解决这个问题,需要一种更加有效的方式来挖掘隐含的标记局部相关性信息,帮助标记分布学习模型的训练。本文提出一种基于样本稀疏表达的标记分布学习(LDL-SR)算法。通过对特征空间中的稀疏模型进行优化,构建样本稀疏系数。在此过程中,获取数据点的低维结构,继而挖掘在高维空间难以发现的低维子空间中的局部相关性信息。通过稀疏系数生成标记分布,从而将隐含在特征空间中的相关性引入标记空间中,有效帮助模型对实例标记分布进行预测。在 11 个数据集上的验证表明,本文算法具有较优的标记分布预测精度和稳定性。

1 标记分布学习

标记分布学习为每一个实例 $\boldsymbol{x}_i$ 构建一个标记分布 $\boldsymbol{D}_i$,其分量 d_i^j 表示该分量对应标记 y_j 对实例 $\boldsymbol{x}_i$ 的描述程度,$d_i^j \in [0,1]$,$\sum_j d_i^j = 1$,即实例能够由标记分布中的标记完全描述。图 1 为风景图实例及其标记分布。单标记学习和多标记学习是标记分布学习的一种特例,反之,标记分布学习是对两者在学习的灵活性上的拓展。更高的灵活性对应更大的

输出空间[12]。例如，对一个标记集合中有 L 个不同标记的学习任务，单标记学习模型有 L 个可能的输出，多标记学习模型有 2^L-1 个可能的输出，而标记分布学习模型的输出空间维度为无穷大。

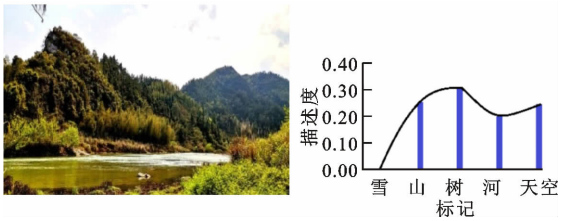


图 1　风景图实例及其标记分布

标记分布学习的目标是学习一个从特征空间到一组由有限标记组成的标记分布的映射。使用 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbf{R}^{n\times t}$ 表示特征空间，每个实例维度为 t。有限标记集为 $Y=\{y_1,y_2,\cdots,y_L\}$，每一个实例 $\boldsymbol{x}_i$ 对应的真实标记分布为 $\boldsymbol{D}_i=[d_i^1,d_i^2,\cdots,d_i^L]$，其中 L 为每个标记分布中标记的数量，为固定值。虽然 d_i^j 并不表示 y_j 对 $\boldsymbol{x}_i$ 正确描述的概率，但其表达的是在完全描述 $\boldsymbol{x}_i$ 的标记集 Y 中，标记 y_j 描述程度所占的比例，故 LDL 的目标也可以表达为从训练集 $T=\{(\boldsymbol{x}_1,\boldsymbol{D}_1),(\boldsymbol{x}_2,\boldsymbol{D}_2),\cdots,(\boldsymbol{x}_n,\boldsymbol{D}_n)\}$ 中学习得到一个条件概率密度函数，并使得用这个函数为未知的实例生成的标记分布与实际标记分布尽可能相似。

近年来，许多实际问题都能够被 LDL 成功解决。例如：Gao 等应用 LDL 进行面部年龄判断[13-14]；在此基础上，Zhang H 等更加细致地选取合适范围的年龄标记集合，构成不同的标记分布，提高年龄预测精度[15]；Zhou Y 等应用 LDL 进行文本情感及强度判断[9]；Zhang Z 等考虑到人群图像实例中标记类别相近的图像实例具有相似的特征，将人群计数问题转换为标记分布问题进行求解[16]。

现有的标记分布学习算法策略可以被大致分为 3 类[17]。第 1 类策略为问题转换，将 LDL 问题转换为现有学习范式可以解决的问题，如 PT-SVM、PT-Bayes 算法。第 2 类策略为改造算法，其主要思想为将现有传统算法进行改进，以适应标记分布学习范式，如 AA-BP、AA-kNN 算法[2]。第 3 类策略是为解决 LDL 问题设计专用算法。设计一个 LDL 专用算法主要包含 3 个部分：输出模型，目标函数以及优化算法。对于输出模型，与 IIS-LLD、BFGS-LLD 算法[18]一致，标记分布学习专用算法一般使用最大熵模型[19]。对于目标函数，一般选用能够衡量两个分布相似度的函数作为目标函数，如 KL 散度[20]、Jeffery 散度[9]。为了解决优化问题，许多优化算法都被应用于标记分布专用算法之中，例如，IIS-LLD 使用改进迭代尺度法作为优化算法[10]，BFGS-LLD 使用 BFGS 算法优化，Zhou D 等使用 L-BFGS 算法进行优化[10]。

在目前现有的专用算法中，部分算法借助标记相关性信息来提高标记分布学习模型的性能，例如：EDL 算法在全局范围内将两个标记间的皮尔森相关系数作为两个标记的相关性[9]；LALOT 算法基于最优传输模型挖掘标记全局相关性[21]；LDLLC 算法引入距离映射函数来计算不同标记间的全局相关性[11]。然而，标记全局相关性无法很好地捕捉潜在的数据特征。本文提出了一种基于局部标记相关性的标记分布学习算法来提高模型预测结果。

2　样本稀疏表达的标记分布学习

本文 LDL-SR 算法属于标记分布学习的专用算法。专用标记分布学习算法主要包含输出模型、目标函数以及优化算法，下面将从此 3 方面介绍本文算法。

与文献[22]相同，本文使用最大熵模型作为输出模型，故标记对实例描述程度的预测值可以表示为

$$\hat{d}_i^k=p(y_k\mid\boldsymbol{x}_i;\boldsymbol{\theta})=\frac{1}{Z_i}\exp\Big(\sum_m\theta_{km}x_i^m\Big)\tag{1}$$

式中：$\hat{d}_i^k$ 为标记 y_k 对实例 $\boldsymbol{x}_i$ 的描述程度预测值；$\boldsymbol{\theta}$ 为待学习的参数，θ_{km} 为 $\boldsymbol{\theta}$ 中的一个元素；x_i^m 为实例 $\boldsymbol{x}_i$ 的第 m 维特征；$Z_i=\sum_k\exp\Big(\sum_m\theta_{km}x_i^m\Big)$ 是归一化项，为了满足 $\sum_k d_i^k=1$ 的性质。

对于实例 $\boldsymbol{x}_i$，由模型生成的标记分布 $\hat{\boldsymbol{D}}_i$ 与真实标记分布 $\boldsymbol{D}_i$ 之间的差距越小越好。文献[23]介绍了许多用于衡量两个分布之间差距的度量算法。文献[12]实验表明，在标记分布学习问题中，KL 散度的表现最为稳定。因此，本文选用 KL 散度作为基础损失函数，公式为

$$A_{\mathrm{KL}}(\boldsymbol{Q}_a,\boldsymbol{Q}_b)=\sum_j Q_a^j\ln\frac{Q_a^j}{Q_b^j}\tag{2}$$

式中 Q_i^j 表示 $\boldsymbol{Q}_i$ 分布的第 j 分量。在 KL 散度度量基础上，为得到模型最佳参数 $\hat{\theta}$，构建目标函数表达式为

$$\min_\theta(E(\theta)+\lambda_1\Omega(\theta)+\lambda_2K(\theta))\tag{3}$$

$$E(\theta)=\sum_{i=1}^n\sum_{j=1}^L d_i^j\ln\Big(\frac{d_i^j}{p(y_j\mid\boldsymbol{x}_i)}\Big)\tag{4}$$

式中 E 表示基础损失函数，该函数基于 KL 散度衡量模型预测值与实际标记分布差距，计算模型在容量为 n 的数据集上预测误差之和。

为防止模型过于复杂从而出现过拟合的现象，使用矩阵的 F 范数实现式(3)的第二项(正则化项)，即

$$\Omega(\boldsymbol{\theta}) = \| \boldsymbol{\theta} \|_{\mathrm{F}}^2 \tag{5}$$

$K(\boldsymbol{\theta})$函数挖掘并利用隐含的标记局部相关性信息。本文使用样本稀疏表达，从特征空间中采集得到数据点隐含的子空间中的相关性信息，并将其转移到标记空间，使标记空间满足相似的稀疏结构。因此，式(3)中的第 3 项可以表达为

$$K(\boldsymbol{\theta}) = \| \boldsymbol{\phi} - \boldsymbol{\phi}\hat{\boldsymbol{S}} \|_{\mathrm{F}}^2 = \| (\boldsymbol{I} - \hat{\boldsymbol{S}}^{\mathrm{T}}) \|_{\mathrm{F}}^2 \tag{6}$$

式中：$\boldsymbol{\phi}=[\hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2, \cdots, \hat{\boldsymbol{D}}_n]$为模型预测的标记分布结果；$\hat{\boldsymbol{S}} \in \mathbf{R}^{n \times n}$为特征空间样本点的稀疏系数，通过限制 $\boldsymbol{\phi}$ 与 $\boldsymbol{\phi}\hat{\boldsymbol{S}}$ 之间的距离来辅助模型进行标记分布的预测。对于 $\hat{\boldsymbol{S}}$，考虑带约束的最小化问题如下

$$\left.\begin{aligned} &\min_{\boldsymbol{S}} \| \boldsymbol{S} \|_1 \\ &\text{s.t. } \boldsymbol{X} = \boldsymbol{X}\boldsymbol{S},\ \mathrm{diag}(\boldsymbol{S}) = \boldsymbol{0} \end{aligned}\right\} \tag{7}$$

稀疏性由矩阵中的非 0 元素数描述，应当以 $\boldsymbol{S}$ 的 l_0 范数作为最小化目标，但是由于其离散的性质，求解成为一个 NP-hard 问题，故将 l_0 范数凸松弛为 l_1 范数，此时依然能得到较理想的稀疏结构[24]。$\mathrm{diag}(\boldsymbol{S})$为 $\boldsymbol{S}$ 的对角元素组成的向量，为了避免 $\boldsymbol{X}$ 中每个数据仅用自己表示，增加约束 $\mathrm{diag}(\boldsymbol{S})=\boldsymbol{0}$。然而，在实际问题中，情况更为复杂。由于数据存在噪声，并且数据点一般都位于仿射子空间而非线性子空间中，故将式(7)重新写为

$$\left.\begin{aligned} &\min_{\boldsymbol{S},\boldsymbol{Z}} (\| \boldsymbol{S} \|_1 + \lambda_z \| \boldsymbol{Z} \|_{\mathrm{F}}^2) \\ &\text{s.t. } \boldsymbol{X} = \boldsymbol{X}\boldsymbol{S} + \boldsymbol{Z}, \boldsymbol{S}^{\mathrm{T}}\boldsymbol{1} = \boldsymbol{1}, \mathrm{diag}(\boldsymbol{S}) = \boldsymbol{0} \end{aligned}\right\} \tag{8}$$

式中：$\boldsymbol{Z}$ 表示特征空间数据中的噪声矩阵；λ_z 为噪声系数，用于平衡式中两项的影响程度。使式(8)取得最优解的矩阵 $\boldsymbol{S}$ 即为本文寻找的稀疏系数 $\hat{\boldsymbol{S}}$。

对于式(8)的凸优化问题，本文使用交替方向乘子法[25]进行求解。为了使目标函数对于每一个变量严格保持凸函数性质，增加两个惩罚项，并且引入辅助变量 $\boldsymbol{A} \in \mathbf{R}^{n \times n}$。此时，式(8)转换为另一个同解的形式

$$\left.\begin{aligned} &\min_{\boldsymbol{S},\boldsymbol{A}} \| \boldsymbol{S} \|_1 + \lambda_z \| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{A} \|_{\mathrm{F}}^2 + \\ &\frac{\rho}{2} \| \boldsymbol{A}^{\mathrm{T}}\boldsymbol{1} - \boldsymbol{1} \|_2^2 + \frac{\rho}{2} \| \boldsymbol{A} - (\boldsymbol{S} - \mathrm{diag}(\boldsymbol{S})) \|_{\mathrm{F}}^2 \\ &\text{s.t. } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{A} = \boldsymbol{S} - \mathrm{diag}(\boldsymbol{S}) \end{aligned}\right\} \tag{9}$$

式中 ρ 为惩罚项的系数。分别为两个等式限制项引入两个拉格朗日乘子 $\boldsymbol{\delta}$、$\boldsymbol{\Delta}$，得到拉格朗日方程

$$\begin{aligned} L(\boldsymbol{S},\boldsymbol{A},\boldsymbol{\delta},\boldsymbol{\Delta}) = {} & \| \boldsymbol{S} \|_1 + \lambda_z \| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{A} \|_{\mathrm{F}}^2 + \\ & \frac{\rho}{2} \| \boldsymbol{A}^{\mathrm{T}}\boldsymbol{1} - \boldsymbol{1} \|_2^2 + \frac{\rho}{2} \| \boldsymbol{A} - (\boldsymbol{S} - \mathrm{diag}(\boldsymbol{S})) \|_{\mathrm{F}}^2 + \\ & \boldsymbol{\delta}^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{1} - \boldsymbol{1}) + \mathrm{tr}(\boldsymbol{\Delta}^{\mathrm{T}}(\boldsymbol{A} - \boldsymbol{S} + \mathrm{diag}(\boldsymbol{S}))) \end{aligned} \tag{10}$$

使用迭代的方式获得最优的 $\boldsymbol{S}$。每一次的优化的步骤如下：先固定 $\boldsymbol{A}$ 和两个拉格朗日乘子迭代优化 $\boldsymbol{S}$；再固定 $\boldsymbol{S}$ 来迭代优化 $\boldsymbol{A}$；最后使用梯度上升的方式更新两个拉格朗日乘子。重复执行这 3 个步骤，直到收敛或达到最大迭代数。

在得到 $\hat{\boldsymbol{S}}$ 后，就可以对本文最终的目标函数进行求解。将式(4)～(6)代入式(3)，得到求解最优模型参数 $\hat{\boldsymbol{\theta}}$ 的目标函数为

$$\begin{aligned} T(\theta) = {} & \sum_i \sum_j d_i^j \ln\left(\frac{d_i^j}{p(y_j \mid \boldsymbol{x}_j)}\right) + \\ & \lambda_1 \| \boldsymbol{\theta} \|_{\mathrm{F}}^2 \prod \lambda_2 \| (\boldsymbol{I} - \hat{\boldsymbol{S}}^{\mathrm{T}})\boldsymbol{\phi}^{\mathrm{T}} \|_{\mathrm{F}}^2 = \\ & \sum_i \sum_j d_i^j \ln d_i^j - \sum_i \sum_j d_i^j \sum_m \theta_{km} x_i^m + \\ & \sum_i \ln \sum_k \exp\left(\sum_m \theta_{km} x_i^m\right) + \\ & \lambda_1 \| \boldsymbol{\theta} \|_{\mathrm{F}}^2 + \lambda_2 \mathrm{tr}(\boldsymbol{\phi}(\boldsymbol{I} - \hat{\boldsymbol{S}})(\boldsymbol{I} - \hat{\boldsymbol{S}}^{\mathrm{T}})\boldsymbol{\phi}^{\mathrm{T}}) \end{aligned} \tag{11}$$

针对目标函数式(11)，本文选用有限内存拟牛顿法(L-BFGS)[26]进行求解。L-BFGS 的主要思想是使用一个迭代更新的矩阵来替代求解海塞矩阵的逆，避免了大量的计算和计算时需要存储整个矩阵，从而能够更高效地求得最优参数 $\hat{\boldsymbol{\theta}}$。$T(\boldsymbol{\theta})$在当前参数 $\boldsymbol{\theta}$ 处的二阶泰勒展开式为

$$T(\boldsymbol{\theta}^{(l+1)}) \approx T(\boldsymbol{\theta}^{(l)}) + \nabla T^{\mathrm{T}}(\boldsymbol{\theta}^{(l)})\boldsymbol{\Delta} + \frac{1}{2}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{\theta}^{(l)})\boldsymbol{\Delta} \tag{12}$$

式中：$\nabla T(\boldsymbol{\theta}^{(l)})$和 $\boldsymbol{H}(\boldsymbol{\theta}^{(l)})$分别为 $T(\boldsymbol{\theta}^{(l+1)})$在参数 $\boldsymbol{\theta}^{(l)}$ 处的梯度和海塞矩阵；$\boldsymbol{\Delta} = \boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^{(l)}$ 为参数的更新变化量。假设 $T(\theta)$为凸函数，其局部极值点为全局极值点，解得使式(12)取得最小值的 $\boldsymbol{\Delta}^{(l)}$ 为

$$\boldsymbol{\Delta}^{(l)} = -\boldsymbol{H}^{-1}(\boldsymbol{\theta}^{(l)})\nabla T(\boldsymbol{\theta}^{(l)}) \tag{13}$$

将 $\boldsymbol{\Delta}^{(l)}$ 作为搜索方向，选择步长 α 后，得到参数的更新公式为

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + \alpha^{(l)}\boldsymbol{\Delta}^{(l)} \tag{14}$$

式中 $\alpha^{(l)}$ 由满足强 Wolfe 条件的线搜索得到[26]。强 Wolfe 条件为

$$T(\boldsymbol{\theta}^{(l)} + \alpha^{(l)}\boldsymbol{\Delta}^{(l)}) \leqslant T(\boldsymbol{\theta}^{(l)}) + c_1\alpha^{(l)}\nabla T^{\mathrm{T}}(\boldsymbol{\theta}^{(l)})\boldsymbol{\Delta}^{(l)} \tag{15}$$

$$|\nabla T(\boldsymbol{\theta}^{(l)} + \alpha^{(l)}\boldsymbol{\Delta}^{(l)})| \leqslant c_2 |\nabla T^{\mathrm{T}}(\boldsymbol{\theta}^{(l)})\boldsymbol{\Delta}^{(l)}| \tag{16}$$

式中 $0<c_1<c_2<1$。

根据式(13)(14)，为了完成优化过程，首先需要计算目标函数 $T(\boldsymbol{\theta})$的梯度$\nabla T(\boldsymbol{\theta})$，梯度公式为

$$\frac{\partial T(\boldsymbol{\theta})}{\partial \theta_{km}}=\sum_i \frac{\exp(\sum_m \theta_{km}\boldsymbol{x}_i^m)\boldsymbol{x}_i^m}{\sum_k \exp(\sum_m \theta_{km}\boldsymbol{x}_i^m)}-\sum_i d_i^i \boldsymbol{x}_i^m+ 2\lambda_1\theta_{km}+2\lambda_2\boldsymbol{\phi}(\boldsymbol{I}-\hat{\boldsymbol{S}}-\hat{\boldsymbol{S}}^{\mathrm{T}}+\hat{\boldsymbol{S}}\hat{\boldsymbol{S}}^{\mathrm{T}})\boldsymbol{\Gamma} \tag{17}$$

$$\boldsymbol{\Gamma}=\left[\frac{\partial \hat{\boldsymbol{D}}_1}{\partial \theta_{km}},\frac{\partial \hat{\boldsymbol{D}}_2}{\partial \theta_{km}},\cdots,\frac{\partial \hat{\boldsymbol{D}}_n}{\partial \theta_{km}}\right] \tag{18}$$

为了避免计算海塞矩阵的逆并节约内存资源，L-BFGS 通过迭代更新矩阵 $\boldsymbol{B}$ 来近似$-\boldsymbol{H}^{-1}(\boldsymbol{\theta}^{(l)})$，迭代公式为

$$\boldsymbol{B}^{(l+1)}=(\boldsymbol{I}-\boldsymbol{\rho}^{(l)}\boldsymbol{s}^{(l)}(\boldsymbol{u}^{(l)})^{\mathrm{T}})\boldsymbol{B}^{(l)}(\boldsymbol{I}-\boldsymbol{\rho}^{(l)}\boldsymbol{u}^{(l)}(\boldsymbol{s}^{(l)})^{\mathrm{T}})+\boldsymbol{\rho}^{(l)}\boldsymbol{s}^{(l)}(\boldsymbol{s}^{(l)})^{\mathrm{T}} \tag{19}$$

$$\boldsymbol{s}^{(l)}=\boldsymbol{\theta}^{(l+1)}-\boldsymbol{\theta}^{(l)} \tag{20}$$

$$\boldsymbol{u}^{(l)}=\nabla T(\boldsymbol{\theta}^{(l+1)})-\nabla T(\boldsymbol{\theta}^{(l)}) \tag{21}$$

$$\boldsymbol{\rho}^{(l)}=\frac{1}{\boldsymbol{s}^{(l)}\boldsymbol{u}^{(l)}} \tag{22}$$

综合本节所述，本文 LDL-SR 算法的伪代码如下。

输入：训练集 $T=\{\boldsymbol{X},\boldsymbol{D}\}$及初始化参数 λ_1、λ_2、λ_Z

输出：标记分布预测值 $\hat{\boldsymbol{D}}_i$

```
由式(10)求得 Ŝ
l←0
初始化模型参数 θ^(l)、海塞阵逆的估计矩阵 B^(l)
由式(17)计算∇T(θ^(l))
repeat
    计算搜索方向 ρ^(l)←−B^(l)∇T(θ^(l))
    由满足式(15)(16)的线搜索计算步长 α^(l)
    更新参数 θ^(l+1)=θ^(l)+α^(l)Δ^(l)
    通过式(17)计算∇T(θ^(l+1))
    由式(20)～(22)分别计算 s^(l)、u^(l)、ρ^(l)
    由式(19)更新估计矩阵 B^(l+1)
    l←l+1
until 满足收敛条件或达到最大迭代数
根据式(1)计算得标记分布 D̂_i，结束
```

3　实　验

在 11 个真实数据集上，使用 5 种评价指标将本文算法与 7 种现有标记分布学习算法进行比较。

3.1　数据集

本文实验所用的 11 个数据集中，前 8 个来自于一组啤酒酵母生长生物实验中的真实数据，其中每个数据集都是一次真实实验的结果。此组数据集原有 10 个，由于本文主要研究借助标记相关性训练标记分布模型，所以去除了标记分布中标记数小于 3 的缺少标记间相关性的数据集。每个酵母基因数据集中包含2 465个基因实例，每个基因由一个长度为 24 的关联动植物轮廓向量描述，标记由生物实验中的离散时间点组成，用不同时间点的基因表达水平(已经过标准化)表示标记对实例的描述程度。SJAFFE、SBU-3DFE 这两个数据集由应用十分广泛的面部表情图像数据集 JAFFE、BU_3DFE 改造而来，均为标记分布数据集。SJAFFE、SBU-3DFE 数据集的每幅图像都由 6 个基本表情构成的标记分布表示，每个表情对某张图像的描述度是 60 个志愿者 1～5 打分后标准化的平均值。HumanGene 数据集是一个包含30 542个样本的大规模数据集，它来自于研究人类基因与疾病之间关系的真实生物实验数据，其标签空间由 68 种不同疾病组成。数据集的详细数据统计信息如表 1 所示。

表 1　数据集详细信息

数据集	n	t	L
Yeast-alpha	2 465	24	18
Yeast-cdc	2 465	24	15
Yeast-cold	2 465	24	14
Yeast-diau	2 465	24	7
Yeast-dtt	2 465	24	6
Yeast-elu	2 465	24	6
Yeast-heat	2 465	24	4
Yeast-spo	2 465	24	4
SJAFFE	213	246	6
SBU-3DFE	2 500	243	6
HumanGene	30 542	36	68

3.2　评价指标

使用 5 种评价指标来对 LDL-SR 的效果进行评价：Sorensen 距离(S)、欧氏距离(E')和相对熵(K)用于衡量两个向量之间的距离，指标越低越优；Fidelity(F)和 Intersection(I)衡量两个向量之间的相似度，指标越高越优。S、E'、K、F、I 的公式分别为

$$S=\frac{\sum_{j=1}^{L}|P_j-Q_j|}{\sum_{j=1}^{c}|P_j+Q_j|};\quad E'=\sqrt{\sum_{j=1}^{L}(P_j-Q_j)^2};$$

$$K=\sum_{j=1}^{L}P_j\ln\frac{P_j}{Q_j};\quad F=\sum_{j=1}^{L}\sqrt{P_jQ_j};$$

$$I=\sum_{j=1}^{L}\min(P_j,Q_j)$$

式中：P_j 表示真实的标记分布 $\boldsymbol{P}$ 的第 j 个分量；Q_j 表示模型预测的标记分布 $\boldsymbol{Q}$ 的第 j 个分量。

3.3 实验设置

为探究本文算法的性能，将 LDL-SR 与 7 种常用的 LDL 算法进行对比，分别是：一种问题转换算法 PT-SVM[22]，一种算法适应性算法 AA-kNN[2]，5 种专用算法 IIS-LLD[2]、EDL[9]、LDLLC[11]、LDL-LCLR[27]、LDLSF[8]。

对于 PT-SVM，使用 libsvm 中的 C-SVC 算法，核函数选用 RBF 核（$c=1.0$，$\gamma=0.01$）；对于 AA-kNN，近邻数 k 设置为 5；对于 LDL-LLC，参数 λ_1 和 λ_2 分别设置为 0.1 和 0.01；对于 IIS-LLD、EDL、LDL-LCLR 及 LDLSF，使用其所在文献中建议的参数值；对于 LDL-SR，将每个数据集随机分为 80% 的训练集和 20% 的测试集进行 10 次实验，参数 λ_1、λ_2 在 $10^{-5}\sim10^{-1}$ 范围内进行调整，选取平均效果最优的值作为参数值，λ_Z 设置为 0.01。

3.4 实验结果

在每个数据集上，使用留出法进行实验，将数据集随机划分为 10 份，每次留出 2 份作为测试集，其他作为训练集。每个算法在不同数据集上将进行 10 次实验，将其在每个数据集上指标的平均值及标准差记录下来，进行对比。使用“平均值±标准差（排名）”格式进行结果展示，其中排名是该算法与其他算法在同一数据集上效果的排序，1 表示效果最好。平均排名表示该算法在 10 个数据集上排名结果的平均值，保留两位小数。实验结果如表 2～6 所示，表中加粗数据为最优值。由于篇幅限制，仅展示部分实验结果，完整结果可扫描本文首页 OSID 码查看。

表 2 E' 指标结果比较

数据集	PT-SVM	AA-kNN	IIS-LLD	LDLLC	LDLSF	LDL-SR
Yeast-cdc	0.029 8±0.000 7(7)	0.030 1±0.000 9(8)	0.029 0±0.001 0(6)	**0.027 6±0.000 9**(1)	0.027 9±0.000 6(3)	0.027 7±0.000 6(2)
Yeast-elu	0.029 3±0.000 8(6)	0.029 7±0.001 0(7)	0.030 7±0.000 9(8)	0.027 7±0.000 6(2)	0.027 9±0.000 3(3)	**0.027 4±0.000 5**(1)
Yeast-diau	0.062 8±0.003 7(8)	0.056 7±0.001 9(6)	0.053 9±0.003 1(2)	0.054 1±0.002 2(3)	0.054 8±0.000 9(4)	**0.053 9±0.000 8**(1)
Yeast-heat	0.062 5±0.002 3(6)	0.062 4±0.002 0(5)	0.070 3±0.003 6(8)	0.060 5±0.001 5(4)	0.059 4±0.000 8(2)	**0.058 6±0.001 6**(1)
Yeast-cold	0.075 3±0.008 0(6)	0.072 4±0.002 7(5)	0.076 7±0.000 4(7)	0.067 9±0.000 3(3)	0.067 7±0.001 8(2)	**0.067 0±0.002 2**(1)
Yeast-dtt	0.051 6±0.002 9(7)	0.050 9±0.002 2(5)	0.053 5±0.002 3(8)	0.047 7±0.001 6(3)	0.048 0±0.001 3(4)	**0.047 4±0.001 2**(1)
SJAFFE	0.161 5±0.030 1(8)	0.123 3±0.018 9(4)	0.150 9±0.014 9(6)	0.157 2±0.009 7(7)	0.114 3±0.007 4(2)	**0.112 2±0.007 7**(1)
HumanGene	0.088 6±0.019 2(5)	0.103 6±0.004 4(7)	0.087 7±0.007 8(4)	0.103 5±0.001 3(8)	**0.085 5±0.002 0**(1)	0.086 1±0.001 5(2)
平均排名	6.82	6.18	6.09	3.64	2.73	1.45

表 3 S 指标结果比较

数据集	PT-SVM	AA-kNN	IIS-LLD	LDLLC	LDLSF	LDL-SR
Yeast-cdc	0.045 8±0.0012(7)	0.046 2±0.001 3(8)	0.044 5±0.001 5(6)	0.042 2±0.001 3(2)	0.042 6±0.000 8(3)	**0.042 2±0.000 7**(1)
Yeast-elu	0.043 8±0.0012(6)	0.044 3±0.001 4(7)	0.047 2±0.001 4(8)	0.041 2±0.000 6(2)	0.041 4±0.000 4(3)	**0.040 8±0.000 8**(1)
Yeast-diau	0.068 6±0.0041(8)	0.062 2±0.002 2(6)	0.059 3±0.003 2(2)	0.059 6±0.002 6(3)	0.060 3±0.001 0(4)	**0.059 3±0.001 0**(1)
Yeast-heat	0.062 7±0.0022(5)	0.063 2±0.001 8(6)	0.069 2±0.003 3(8)	0.061 0±0.001 6(4)	0.060 0±0.000 8(2)	**0.059 2±0.000 8**(1)
Yeast-cold	0.065 4±0.0069(7)	0.063 0±0.002 4(4)	0.065 3±0.003 4(6)	0.058 9±0.001 9(3)	0.058 6±0.001 6(2)	**0.058 1±0.002 0**(1)
Yeast-dtt	0.044 7±0.0024(7)	0.044 3±0.001 7(6)	0.048 0±0.002 3(8)	0.041 5±0.001 3(3)	0.041 8±0.001 1(4)	**0.041 2±0.000 7**(1)
SJAFFE	0.154 5±0.0253(7)	0.120 1±0.014 8(4)	0.148 2±0.013 0(6)	0.155 1±0.008 2(8)	0.113 3±0.007 2(2)	**0.110 0±0.007 1**(1)
HumanGene	0.219 1±0.0124(6)	0.254 8±0.003 6(7)	0.218 7±0.005 4(4)	0.272 4±0.002 6(8)	**0.213 5±0.003 2**(1)	0.217 6±0.001 5(2)
平均排名	6.82	6.18	5.82	3.82	2.73	1.36

表 4 K 指标结果比较

数据集	PT-SVM	AA-kNN	IIS-LLD	LDLLC	LDLSF	LDL-SR
Yeast-cdc	0.007 6±0.000 4(7)	0.007 9±0.000 4(8)	0.007 2±0.000 2(5)	**0.006 8±0.000 1**(1)	0.007 0±0.000 3(3)	0.007 1±0.000 3(4)
Yeast-elu	0.006 8±0.000 5(6)	0.007 1±0.000 6(8)	0.007 1±0.000 4(7)	0.006 8±0.000 3(5)	0.006 3±0.000 1(2)	**0.006 2±0.000 2**(1)
Yeast-diau	0.016 7±0.001 7(8)	0.014 5±0.001 0(6)	0.014 1±0.001 3(5)	**0.013 0±0.001 0**(1)	0.013 3±0.000 5(2)	0.013 5±0.000 5(3)
Yeast-heat	0.014 1±0.001 0(5)	0.014 1±0.001 0(5)	0.018 2±0.001 6(8)	0.013 1±0.000 6(4)	**0.012 7±0.000 4**(1)	0.012 8±0.000 8(2)
Yeast-cold	0.014 6±0.003 3(6)	0.013 6±0.001 1(5)	0.015 5±0.001 5(8)	0.012 0±0.000 6(3)	0.012 0±0.000 7(4)	0.012 0±0.000 5(2)
Yeast-dtt	0.007 1±0.000 9(8)	0.006 9±0.000 7(6)	0.006 8±0.000 5(5)	0.006 1±0.000 5(2)	0.006 2±0.000 3(3)	**0.006 1±0.000 3**(1)
SJAFFE	0.076 4±0.022 0(8)	0.052 4±0.026 3(3)	0.071 2±0.013 5(7)	0.075 4±0.008 3(6)	0.055 1±0.002 1(4)	**0.041 5±0.004 0**(1)
HumanGene	0.246 6±0.032 4(6)	0.289 4±0.017 6(7)	0.241 4±0.031 0(4)	0.326 2±0.007 7(8)	**0.238 0±0.002 2**(1)	0.239 4±0.001 7(2)
平均排名	6.95	6.50	5.91	3.50	2.73	1.95

表 5　F 指标结果比较

数据集	PT-SVM	AA-kNN	IIS-LLD	LDLLC	LDLSF	LDL-SR
Yeast-cdc	0.9980±0.0001(7)	0.9980±0.0001(7)	0.9982±0.0012(6)	0.9983±0.0002(2)	0.9982±0.0000(3)	**0.9983±0.0001**(1)
Yeast-elu	0.9983±0.0002(6)	0.9982±0.0002(7)	0.9982±0.0035(8)	0.9984±0.0001(3)	0.9984±0.0001(3)	**0.9985±0.0001**(1)
Yeast-diau	0.9957±0.0004(8)	0.9963±0.0003(5)	0.9964±0.0036(4)	0.9962±0.0002(6)	0.9965±0.0001(2)	**0.9966±0.0001**(1)
Yeast-heat	0.9964±0.0003(5)	0.9964±0.0003(5)	0.9954±0.0042(8)	0.9966±0.0003(4)	0.9968±0.0001(2)	**0.9969±0.0002**(1)
Yeast-cold	0.9963±0.0008(6)	0.9966±0.0003(5)	0.9960±0.0039(8)	0.9967±0.0023(4)	0.9969±0.0002(2)	**0.9970±0.0001**(1)
Yeast-dtt	0.9982±0.0003(7)	0.9982±0.0002(6)	0.9983±0.0013(5)	0.9985±0.0002(2)	**0.9985±0.0001**(1)	0.9984±0.0001(3)
SJAFFE	0.9815±0.0050(8)	0.9868±0.0050(4)	0.9827±0.0030(6)	0.9818±0.0018(7)	0.9884±0.0018(2)	**0.9897±0.0009**(1)
HumanGene	0.9443±0.0281(6)	0.9324±0.0034(7)	0.9454±0.0049(3)	0.9267±0.0014(8)	**0.9461±0.0009**(1)	0.9455±0.0014(2)
平均排名	6.91	6.36	5.73	4.14	2.32	1.45

表 6　I 指标结果比较

数据集	PT-SVM	AA-kNN	IIS-LLD	LDLLC	LDLSF	LDL-SR
Yeast-cdc	0.955 4±0.001 2(7)	0.953 8±0.001 3(8)	0.955 6±0.001 5(6)	**0.957 7±0.001 3**(1)	0.957 4±0.000 8(3)	0.957 6±0.001 0(2)
Yeast-elu	0.956 2±0.001 2(6)	0.955 7±0.001 4(7)	0.952 8±0.001 5(8)	0.958 0±0.001 3(4)	0.958 6±0.000 4(2)	**0.959 2±0.000 8**(1)
Yeast-diau	0.931 4±0.004 1(8)	0.937 8±0.002 2(6)	0.940 7±0.003 3(2)	0.940 4±0.002 6(3)	0.939 7±0.001 0(4)	**0.940 7±0.001 0**(1)
Yeast-heat	0.937 3±0.002 2(4)	0.936 8±0.001 8(5)	0.930 9±0.003 3(7)	0.938 9±0.001 4(4)	0.940 0±0.000 8(2)	**0.940 8±0.001 4**(1)
Yeast-cold	0.934 6±0.006 9(7)	0.937 0±0.002 4(5)	0.934 7±0.003 4(6)	0.941 4±0.001 9(3)	0.941 4±0.001 6(2)	**0.941 9±0.002 0**(1)
Yeast-dtt	0.955 3±0.002 4(7)	0.955 7±0.001 7(6)	0.952 0±0.002 3(8)	0.958 5±0.001 3(2)	0.958 2±0.001 1(4)	**0.958 8±0.001 9**(1)
SJAFFE	0.845 5±0.025 3(7)	0.879 8±0.014 8(4)	0.851 8±0.013 0(6)	0.844 9±0.008 2(8)	0.886 7±0.007 2(2)	**0.890 0±0.006 1**(1)
HumanGene	0.781 0±0.012 6(6)	0.745 1±0.003 6(7)	0.781 3±0.005 4(4)	0.727 6±0.002 6(8)	**0.785 0±0.001 2**(1)	0.784 3±0.001 5(2)
平均排名	6.73	6.18	5.73	3.91	2.64	1.36

通过对表 2～6 进行分析，可以得出以下结论。

(1)在 11 个数据集中，本文 LDL-SR 算法解决 LDL 问题的效果在多数情况下优于其他 7 种对比算法的，LDL-SR 在不同种类的数据集上均能取得稳定的表现。

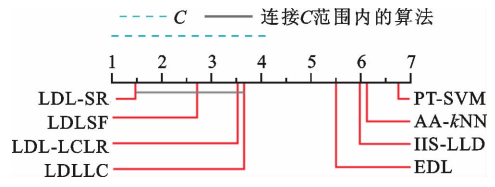
(2)使用了标记间相关性的 EDL、LDLLC 和 LDL-SR 算法的效果要好于其他算法的。这是因为标记相关性能够为标记分布中的标记建立联系，为标记分布模型预测提供额外信息。

(3)LDL-SR 算法使用样本局部标记相关性，效果要明显优于将标记相关性视为全局的 EDL 和 LDLLC 算法的，证明了在样本局部范围内采集标记相关性的可行性。同时也说明在现实数据集中，若利用全局相关性将所有样本点视为同一组，共享标记相关性，会因相关性粒度过大而误导模型预测。局部相关性得益于小粒度的相关性范围，更加适合于解决实际问题。

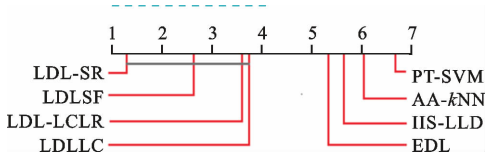
(4)LDLSF 算法注重特征的选取，所以在特征维度较高的数据集上能够取得较好的效果，但是在中低维度的特征空间中则无法很好地发挥作用。相比之下，本文 LDL-SR 算法在不同特征空间维度的数据集中都能取得较理想的效果。

(5)LDL 专用算法的效果要优于问题转换算法和改造算法的。这是因为专用算法在设计时就以最大化标记分布预测值与实际值之间的相似度为目标，而问题转换算法将标记分布训练集转换为单标记训练集，改造算法如 AA-kNN 将标记分布中的描述度分量作为独立值进行处理，导致了对训练有益信息的丢失。

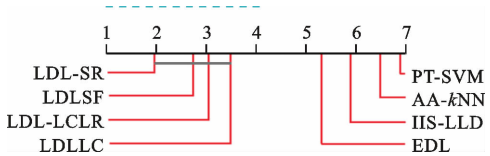
Nemenyi 检验是在各总体分布位置不完全相同的情况下，需要进一步判断是哪两个总体分布位置不同时常用的两两比较检验法。在 Nemenyi 检验中，若两个算法在某一评价指标下的综合排序结果小于临界值阈 C，则说明两者效果无显著性差异，否则两者效果差异显著。本文设定 C 为3.165 8。为了更好地对比算法之间效果的差异，本文使用显著性水平为 5% 的 Nemenyi 检验进行算法差异性比较，结果如图 2 所示。可以看出，本文算法与大部分算法间具有显著性差异。



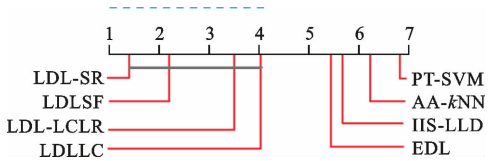
(a)E'



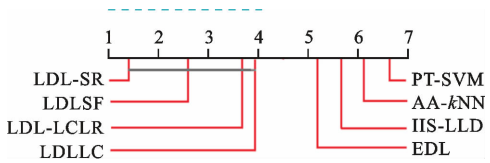
(b)S



(c)K



(d)F

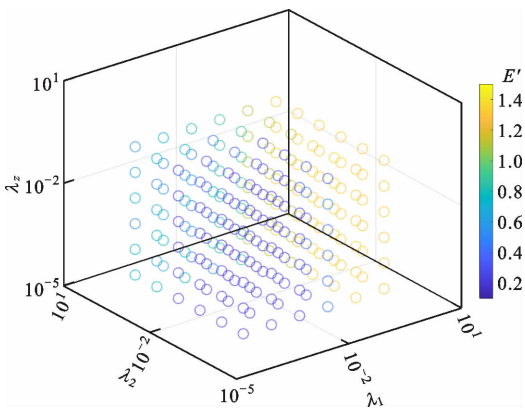


(e)I

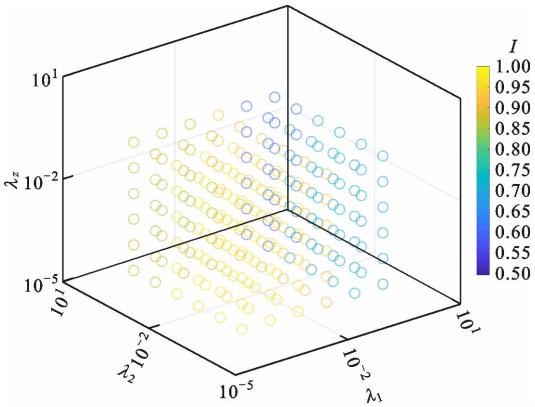
图 2 Nemenyi 检验在 8 个标记分布学习算法上的临界差异

3.5 参数分析

本小节对 LDL-SR 中的参数对算法结果的影响进行研究。由于篇幅限制，仅展示在 Yeast-alpha 数据集上使用 E' 和 I 评价指标的分析结果，如图 3 所示。可以看出：仅变动参数 λ_z 时，结果变化十分微小，故取中间值 0.01；当 λ_1 为 1 或 10 时，正则化项占据主导，结果会变得很差；当 λ_1、λ_2 都变得很大时，目标函数中描述标记分布相似度的 KL 散度无法起到主导作用，也会影响算法效果。因此，LDL-SR 中 λ_1、λ_2 两个参数应该在范围 $10^{-4}\sim10^{-2}$ 内进行选取。从图 3 可以看出，只要在合适的范围内进行参数选择，本文算法都能有稳定的表现，具有较好的鲁棒性。



(a)E'

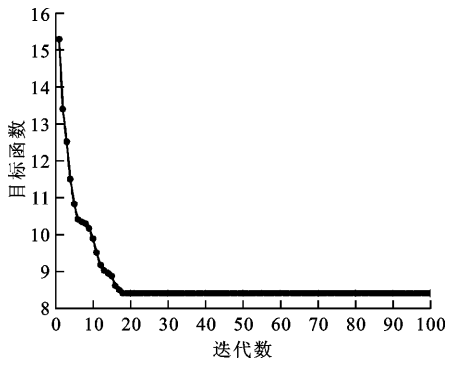


(b)I

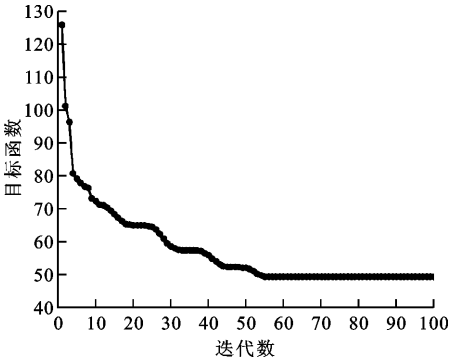
图 3 Yeast-alpha 数据集上参数对结果的影响

3.6 算法收敛性

本小节对 LDL-SR 算法的收敛性进行研究，在数据集 SJAFFE、SBU-3DFE 上进行实验。使用 L-BFGS 优化算法求解 LDL-SR 算法时的收敛性如图 4 所示。可以看出：目标函数随迭代数的增大而降



(a)SJAFFE 数据集



(b)SBU-3DFE 数据集

图 4 SJAFFE 和 SBU-3DFE 数据集上 LDL-SR 算法的收敛性

低，在迭代数一定后趋于定值；对于数据集 SJ-FAAE，LDL-SR 算法在迭代 20 次左右后目标函数基本维持不变，而在 SBU-3DFE 数据集上需要迭代 55 次左右。

4　结　论

标记分布学习是一种能够解决复杂标记多义性的机器学习范式。由于标记分布中通常包含多个标记，因此利用标记间的相关性是一种提高标记分布模型预测能力的有效途径。本文提出一种基于样本稀疏表达的标记分布学习算法，并且得出以下结论。

(1)在实际问题中，标记相关性通常在局部样本空间内共享，并非所有样本共享相同的标记相关性。本文借助稀疏表达模型，挖掘样本点中的局部相关性信息，将稀疏系数引入标记空间来预测隐含的标记相关性，借助标记局部相关性来帮助标记分布模型的训练。

(2)在 11 个数据集上进行的实验结果表明，本文算法较其他对比算法能够取得更高的预测精度，也具有更好的稳定性。同时，本文算法在合理范围内进行参数取值都能取得理想的效果，具有较强的鲁棒性。

(3)本文算法在高维特征空间数据集上还具有提升空间，后续研究中将考虑增加特征选择步骤来提高算法在高维特征空间中的预测精度。

参考文献：

[1] READ J, PFAHRINGER B, HOLMES G. Multi-label classification using ensembles of pruned sets [C]// Proceedings of the 2008 8th IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2008: 995-1000.

[2] GENG X, SMITH-MILES K, ZHOU Z H. Facial age estimation by learning from label distributions [C]// Proceedings of the 24th AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2010: 451-456.

[3] GENG Xin, YIN Chao, ZHOU Zhihua. Facial age estimation by learning from label distributions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2401-2412.

[4] GENG Xin. Label distribution learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1734-1748.

[5] WANG J, GENG X. Theoretical analysis of label distribution learning [C]// Peoceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2019: 5256-5263.

[6] WANG Jing, GENG Xin. Classification with label distribution learning [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California, USA: International Joint Conferences on Artificial Intelligence Organization, 2019: 3712-3718.

[7] XU Suping, SHANG Lin, SHEN Furao. Latent semantics encoding for label distribution learning [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California, USA: International Joint Conferences on Artificial Intelligence Organization, 2019: 3982-3988.

[8] REN Tingting, JIA Xiuyi, LI Weiwei, et al. Label distribution learning with label-specific features [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California, USA: International Joint Conferences on Artificial Intelligence Organization, 2019: 3318-3324.

[9] ZHOU Ying, XUE Hui, GENG Xin. Emotion distribution recognition from facial expressions [C]// Proceedings of the 23rd ACM International Conference on Multimedia. New York, USA: ACM, 2015: 1247-1250.

[10] ZHOU Deyu, ZHANG Xuan, ZHOU Yin, et al. Emotion distribution learning from texts [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 638-647.

[11] JIA X, LI W, LIU J, et al. Label distribution learning by exploiting label correlations [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2018: 3310-3317.

[12] 赵权，耿新. 标记分布学习中目标函数的选择 [J]. 计算机科学与探索，2017，11(5)：708-719.
ZHAO Quan, GENG Xin. Selection of target function in label distribution learning [J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(5): 708-719.

[13] GAO Binbin, ZHOU Hongyu, WU Jianxin, et al. Age estimation using expectation of label distribution learning [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. California, USA: International Joint Conferences on Artificial Intelligence Organization, 2018: 712-718.

[14] GAO Binbin, XING Chao, XIE Chenwei, et al. Deep label distribution learning with label ambiguity [J].

IEEE Transactions on Image Processing, 2017, 26(6): 2825-2838.

[15] ZHANG H Y, ZHANG Y, GENG X, Practical age estimation using deep label distribution learning [J/OL]. Frontiers of Computer Science. [2020-04-10]. https://doi.org/10.1007/s11704-020-8272-4.

[16] ZHANG Zhaoxiang, WANG Mo, GENG Xin. Crowd counting in public video surveillance by label distribution learning [J]. Neurocomputing, 2015, 166: 151-163.

[17] 耿新, 徐宁. 标记分布学习与标记增强 [J]. 中国科学: 信息科学, 2018, 48(5): 521-530.
GENG Xin, XU Ning. Label distribution learning and label enhancement [J]. Scientia Sinica: Informationis, 2018, 48(5): 521-530.

[18] GENG Xin. Label distribution learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1734-1748.

[19] BERGER A L, DELLA PIETRA V J, DELLA PIETRA S A. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 39-71.

[20] YANG Xu, GAO Binbin, XING Chao, et al. Deep label distribution learning for apparent age estimation [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). Piscataway, NJ, USA: IEEE, 2015: 344-350.

[21] PENG Z, ZHOU Z. Label distribution learning by optimal transport [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2018: 4506-4513.

[22] GENG Xin, WANG Qin, XIA Yu. Facial age estimation by adaptive label distribution learning [C]// Proceedings of the 2014 22nd International Conference on Pattern Recognition. Piscataway, NJ, USA: IEEE, 2014: 4465-4470.

[23] CHA S H. Comprehensive survey on distance/similarity measures between probability density functions [J]. International Journal of Mathematical Models and Methods in Applied Sciences, 2007, 1(4): 300-307.

[24] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: algorithm, theory, and applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.

[25] BOYD S, PARIKH N, CHU E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122.

[26] NOCEDAL J, WRIGHT S J. Numerical optimization [M]. Cham, Germany: Springer, 1999: 176-184.

[27] REN Tingting, JIA Xiuyi, LI Weiwei, et al. Label distribution learning with label correlations via low-rank approximation [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California, USA: International Joint Conferences on Artificial Intelligence Organization, 2019: 3325-3331.

[本刊相关文献链接]

程士卿,郝问裕,李晨,等.低秩张量分解的多视角谱聚类算法.2020,54(3):119-125.[doi:10.7652/xjtuxb202003015]

吴静然,丁恩杰,崔冉,等.采用多尺度注意力机制的旋转机械故障诊断方法.2020,54(2):51-58.[doi:10.7652/xjtuxb202002007]

张志强,孙若斌,徐冠基,等.采用非相关字典学习的滚动轴承故障诊断方法.2019,53(6):29-34.[doi:10.7652/xjtuxb201906005]

余星达,陈文杰,王鼎,等.非接触式身份识别的深度学习算法.2019,53(4):122-127.[doi:10.7652/xjtuxb201904018]

傅翠娇,钱德沛,栾钟治.采用机器学习算法的软件能耗感知模型及其应用.2018,52(12):70-76.[doi:10.7652/xjtuxb201812011]

张西宁,雷威,杨雨薇,等.采用自适应基因粒子群算法优化隐马尔科夫模型的方法及应用.2018,52(8):1-8.[doi:10.7652/xjtuxb201808001]

颛孙少帅,杨俊安,刘辉,等.采用双层强化学习的干扰决策算法.2018,52(2):63-69.[doi:10.7652/xjtuxb201802010]

贾晓琳,樊帅帅,罗雪,等.应用非线性加权的集成学习软件缺陷序列预测算法.2017,51(7):156-161.[doi:10.7652/xjtuxb201707022]

王万召,王杰.采用限定记忆极限学习机的过热汽温逆建模研究.2014,48(2):32-37.[doi:10.7652/xjtuxb201402006]

冯斌,梅雪松,杨军,等.数控机床摩擦误差自适应补偿方法研究.2013,47(11):65-69.[doi:10.7652/xjtuxb201311012]

(编辑 陶晴)