计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700077

基于边际概率分布匹配的主动标记分布学习

董心悦　范瑞东　侯臣平
国防科技大学文理学院　长沙 410008
(dongxinyue96@163.com)

摘　要　标记分布学习是在以标记分布标注的示例上学习的新型学习范式,近年来已成功应用于面部年龄估计、头部姿势估计和情感识别等实际场景中。在标记分布学习中,需要足够多的标记分布数据才能训练出预测性能好的模型。然而,标记分布学习有时会面临标记数据不足和注释成本太高的困境。基于边际概率分布匹配的主动标记分布学习(Active Label Distribution Learning Based on Marginal Probability Distribution Matching,ALDL-MMD)算法是针对标记分布学习注释成本过高的问题而设计的,以减少训练模型所需的标注数据量,从而降低注释成本。ALDL-MMD 算法训练了一个线性回归模型,在保证其训练误差最小的同时,学习一个反映未标记数据上选点需求的稀疏向量,使选点后的训练集和未标记集的数据分布尽量相似,并对这个向量做松弛化处理,以简计算。在多个标记分布数据集上的实验结果表明,在“Canberra Metric”和“Intersection”这两个衡量标记分布的指标上,ALDL-MMD 算法优于已有的主动示例选择方法,体现了其在降低注释成本方面的有效性。

关键词:主动学习;标记分布学习;最大平均差异;边际概率分布匹配;线性模型
中图法分类号　TP391

Active Label Distribution Learning Based on Marginal Probability Distribution Matching

DONG Xin-yue,FAN Rui-dong and HOU Chen-ping
College of Liberal Arts and Sciences,National University of Defense Technology,Changsha 410008,China

Abstract　Label distribution learning (LDL)is a new learning paradigm for learning on instances labeled with label distribution, and has been successfully applied to real world scenes such as face age estimation,head pose estimation,and emotion recognition in recent years. In label distribution learning,enough data labeled by label distribution is needed when people train a model with good prediction performance. However,label distribution learning sometimes faces the dilemma that labeled data is insufficient and that marking enough label distribution data means high annotation cost. The Active label distribution learning based on marginal probability distribution matching (ADLD-MMD)algorithm is designed to solve the problem of high annotation cost for label distribution learning,by reducing the amount of labeled data required to train the model and reducing the annotation cost accordingly. The ALDL-MMD algorithm trains a linear regression model. While ensuring the minimum training error of the linear regression model,it learns a sparse vector that reflects that which instance in the unlabeled data set are selected,so that the data distribution of the training data set and unlabeled data set after instance selection is as similar as possible. We relax the vector for easy calculation. An effective method to optimize the objective function in ALDL-MMD is given,and proof for the convergence of ALDL-MMD is also provided. The experimental results on multiple label distribution data sets show that the ALDL-MMD algorithm is superior to the existing active example selection methods on the two evaluation measures of "Canberra Metric" (distance) and “Intersection” (similarity) to measure that what degree of the label distribution of the instance is accurate,which reflects its effectiveness in reducing annotation costs.

Keywords　Active learning,Label distribution learning,Maximum mean discrepancy,Marginal probability distribution matching,Linear model

标记多义性问题是当前机器学习研究中的热门课题。传统的监督学习建立了从特征空间到标记空间的有效映射。反映标记对数据的描述的学习范式主要分为两种:单标记学习(Single Label Learning,SLL)和多标记学习(Multi-Label

到稿日期:2020-05-05　返修日期:2020-07-13　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61922087,61906201);湖南省杰出青年自然科学基金(2019JJ20020)
This work was supported by the National Natural Science Foundation of China(61922087,61906201) and Natural Science Foundation for Distinguished Young Scholars of Hunan Province (2019JJ20020).
通信作者:侯臣平(hcpnudt@hotmail.com)

Learning，MLL)[1]。SLL 认为所有的训练示例都由一个标记标注，而 MLL 则允许训练示例由多个标记标注。二者都能够解决“示例可以被哪个(些)示例描述”这一基本问题，但都不能解决“每个标记可以在多大程度上描述示例”这个更深层的问题。然而，在很多现实问题中，示例不仅与多个标记相关，而且每个标记的重要程度或者描述示例的程度也各不相同。传统的监督学习方法无法直接解决这个问题。

针对此问题，Geng 等[2]提出了一种新的学习范式，对于一个示例，对每个可能的标记赋予一个实数 d_x^y 来表示该标记对示例的描述度。所有标记的描述度满足 $d_x^y \in [0,1]$ 且 $\sum_y d_x^y = 1$ 的数据结构称为标记分布，在标记分布标注的数据集上学习的过程则称为标记分布学习(Label Distribution Learning，LDL)。近年来，标记分布学习已经在很多现实场景中得到应用，并且取得了良好的效果，如面部年龄估计[3]、头部姿势估计[4]、情感识别[5]和文本挖掘[6]等。标记分布学习需要足够的训练数据才能保证算法的良好效果。与由逻辑数“0”和“1”标注的数据相比，人类注释者在使用标记分布注释示例时，还需要识别每个可能标记的相对重要性，这使得标注标记分布数据需要耗费更高的注释成本，标记分布标注的数据也就更难以获得。

为解决训练数据注释成本较高的问题，经典的学习范式——主动学习通过引入模型来极大地减少训练一个模型所需的标记示例数量，从而降低人类注释的成本。其选点策略的核心要素是衡量未标记数据使模型训练至最优的潜力，并选择最有潜力的数据，查询其标记后将其加入训练集参与模型的训练。现有的主动学习算法可分为两类：一类查询信息量最大的示例，如基于委员会的查询[7-8]、基于不确定性抽样的方法[9]和最佳实验设计[10]；另一类旨在选择最具代表性的示例，如转导实验设计[11]和基于聚类的方法[12]。

尽管主动学习算法在传统机器学习中取得了巨大的成功，但此前没有专门为标记分布学习场景而设计的主动学习算法，而直接强行采用已有的主动学习算法并不能取得令人满意的性能，这可能是因为标记分布与传统逻辑标记的标注形式有很大的不同，并且更为复杂。标记分布的标注方式是无穷的，而 SLL 和 MLL 只是标记分布学习的特例。由于标记分布学习能降低注释成本，且使用较少的训练数据训练模型就能达到良好的效果，因此设计出针对标记分布学习场景的主动学习算法很有必要。

文中提出一种新的主动标记分布算法 ALDL-MMD。这种算法训练一个线性回归模型，在最小化该模型的训练损失的同时，选择数据集中有代表性的数据，使已标注的训练数据与未标注的无标记数据的数据分布在每轮选点中尽可能相似。采用边际概率分布衡量数据分布之间的差别，并通过学习一个稀疏的二进制向量来进行选点。为了简化计算，将其松弛求解。文中给出了有效的方法来优化 ALDL-MMD 中的目标，并且证明了 ALDL-MMD 的收敛性。实验验证了相比于运用在标记分布场景中的传统的主动学习算法，ALDL-MMD 算法用更少的标记数据就能达到更高的预测精度，说明该算法能够降低标记分布学习的注释成本。

1 相关工作

1.1 LDL

LDL[2]是一种监督学习框架，它将在由逻辑标记标注的数据上的学习转换为在由标记分布数据上学习，标记分布表示示例描述中涉及的所有标记的相对重要性。针对 LDL 所设计的算法通常分为 3 类。1)问题转换算法，旨在通过采样将具有标记分布的训练示例转换为加权的单标记示例，以将 LDL 问题转换为 SLL 或 MLL 学习问题。采样后，可以将 SVM[13]和朴素贝叶斯[14]应用于二元分类问题。这类算法包括 PT-SVM 和 PT-Bayes[2]。2)适应算法，如 AA-Bayes 和 AA-BP 算法[2]，将逻辑标记适应于某些特定的算法，将多标记学习算法适应于标记分布数据，从而将传统的 MLL 算法扩展到标记分布算法。3)专用算法，它是直接针对 LDL 问题而设计的，典型代表是 SA-IIS 和 SA-BFGS[2]。

SA-IIS 和 SA-BFGS 都假定含参数的模型 $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ 是最大熵模型，$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{\boldsymbol{Z}}\exp(\sum_k \boldsymbol{\theta}_{y,k} g_k(\boldsymbol{x}))$。其中，$\boldsymbol{Z} = \sum_y \exp(\sum_k \boldsymbol{\theta}_{y,k} g_k(\boldsymbol{x}))$ 是归一化项，$g_k(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的第 k 个特征。目标函数为 $T(\boldsymbol{\theta}) = \sum_{i,j} d_{x_i}^{y_j}$。

当目标函数使用增强迭代缩放(Improved Iterative Scaling，IIS)[15]优化时，称其为 SA-IIS；当 $T'(\boldsymbol{\theta}) = -T(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}^{(l)}$ 的二阶泰勒展开是梯度且海森矩阵为 $T'(\boldsymbol{\theta}^{(l)})$，线性搜索牛顿法使用 $\Delta^{(l)}$ 作为搜索方向 $\boldsymbol{p}^{(l)}$，使用 $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + a^{(l)}\boldsymbol{p}^{(l)}$ 更新参数，搜索步长 $a^{(l)}$ 满足强 Wolfe 条件，使用 BFGS 拟牛顿法[16]用迭代更新的矩阵近似 $\boldsymbol{H}^{(-1)}(\boldsymbol{\theta}^{(l)})$ 时，称其为 SA-BFGS。

1.2 主动学习

主动学习最初是为解决二进制分类问题而创建的一种有效的机器学习范式，由 Simon 于 1974 年首次提出。主动学习算法选择对模型性能最有利的示例。在学习过程中，基于一定的选点准则选择未标记的示例，并将这些示例添加到训练集中以进一步训练模型。选择查询和添加到训练集的不同示例对最终模型有不同的影响。最常见的思路是查询信息量最大或最具代表性的示例[17]。

查询信息量最大的示例的一个代表算法是委员会投票选择算法(Query-By-Committee，QBC)，这是 Seung 基于已有的分类标记数据于 1992 年提出的[7]。具体方法是由两个或更多分类器组成一个“委员会”，委员会成员预测未标注示例的标记，然后选择最不一致的示例进行查询。该算法认为选择的示例是信息量最大的。QBC 选点策略减小了搜索空间，将被查询示例添加到训练集中，并缩减了整个搜索空间，从而加快了学习过程。

查询有代表性的示例的典型方法是流形适应性实验设计(Manifold Adaptive Experimental Design，MAED)[18]。该方法通过使用由数据依赖的范式重现内核希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)[19]，改变 RKHS 的结构以反映数据的结构，然后在流形自适应内核空间执行常规的最佳实验设计。求解式对实验设计进行核处理，从而产生用于文本分类的非线性流形自适应数据选择，其优化目标为：

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\boldsymbol{R}^n}\sum_{i=1}^{n}\left(\|\boldsymbol{x}_i+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_i^2\|+\sum_{j=1}^{n}\frac{\alpha_{i,j}^2}{\beta_j}\right)+\gamma\|\boldsymbol{\beta}\|_1$$

$$\text{s.t. } \beta_i\geqslant 0, i=1,\cdots,n$$

2 基于边际概率分布匹配的主动标记分布学习方法

本节提出了基于边际概率分布匹配的主动标记分布学习算法。首先介绍文中所用符号的含义，接着提出边际概率分布匹配的主动标记分布学习模型，并给出模型的求解方法。

2.1 基本定义

用黑斜体大写表示矩阵，黑斜体小写表示向量，$\|\cdot\|_F$ 表示 Frobenius 范数，具体符号说明如表 1 所列。

表 1 符号说明

Table 1 Notations

符号	含义
L	训练数据集
U	未标注数据集
Q	每轮选点中查询的数据集
n_l	L 中的示例数
n_u	U 中的示例数
b	Q 中的示例数
λ	平衡参数
c	标记数
d	特征维数
y_i	示例 x_i 的标记分布

2.2 ALDL-MMD 算法

本小节提出了标记分布学习场景下的一种新型主动学习算法 ALDL-MMD。这种算法训练一个线性回归模型，在最小化模型的训练损失的同时，选择数据集中具有代表性的数据，使已标注的训练数据与未标注的无标记数据的数据分布尽可能相似。直观的方法就是在每一轮迭代选点中，使加入查询数据后的训练数据与无标记数据集的数据在标记空间的联合概率分布相近。这样的选择最能代表未标记示例分布的查询示例集，使模型在未标记的数据集和来自同一分布的数据上具有良好的泛化性能。假定两个数据集数据在特征空间到标记空间的映射为 $f(\boldsymbol{X})=\boldsymbol{WX}$，是一个线性回归模型，将分布的总体期望之差作为衡量这两个分布相近与否的损失函数。

$$\sum_{i\in L}\|f(\boldsymbol{x}_i)-\boldsymbol{y}_i\|_F^2+\lambda\|E_L(f(\boldsymbol{X}))-E_U(f(\boldsymbol{X}))\|_F^2 \quad (1)$$

由于训练数据和未标注数据都是从同一个原始数据集中随机选择的，可以认为它们所属的基础分布是相同的，从而可以将问题转化为使它们的边际概率相近。假设未标注数据集 U 有 n_u 个未标记数据，训练集 L 有 n_l 个标记数据，每轮选点时，在未标注数据集 U 中选择一个有 b 个示例的数据集 Q 进行查询，使新形成的训练集 $L\cup Q$ 的分布与新形成的未标注集 $U\backslash Q$ 的分布相近，即 $\|E_{L\cup Q}(f(\boldsymbol{X}))-E_{U\backslash Q}(f(\boldsymbol{X}))\|_F^2$ 尽可能小。在这种情况下，需要采用一个指标来有效地衡量数据分布之间的差异。在迁移学习中，只有当测试数据与训练数据属于相同的基础分布时，传统的机器学习算法才能在分类器上提供性能保证。因此，迁移学习中能够有效地衡量训练集与测试集之间数据分布差异的方法和指标同样也适用于本文问题。最大均值差异（Maximum Mean Discrepancy，MMD）已被广泛且成功地用于迁移学习应用中[20]，以确保训练和测试数据之间的边际分布相似。

最大均值差异的基本原理是：假设 S 和 P 是从目标示例群中随机抽取的两组示例，示例数分别为 n_s 和 n_p，令 s 和 p 分别为基于示例集 S 和 P 定义的概率分布。MMD 的基本原理是找到一个对两个不同分布的 s 和 p 具有不同期望的函数 f，以便在对不同分布中的示例进行经验评估时可以评估 S 和 P 的分布 p 和 s 是否相似，即 $\|E_s(f(\boldsymbol{S}))-E_p(f(\boldsymbol{P}))\|_F^2$ 尽可能小。令 f 为函数 $F:S\to R^c$ 的一类，将最大平均差异及其经验估计定义为：如果 F 含有足够多不同类型的映射，则 MMD$[F,Q,P]$将在当且仅当 $p=s$ 时消失，即 $\|E_s(f(\boldsymbol{S}))-E_p(f(\boldsymbol{P}))\|_F^2=0$。在之前的工作中，MMD 是应用于数据层面来判定不同数据集的数据是否来自同一个数据分布。考虑到我们研究的任务是为数据预测出标记分布，那么不妨认为其计算 MMD 的一类函数是从特征空间到标记空间的映射 $f(\boldsymbol{X})=\boldsymbol{WX}$，从而检测 p 和 s 之间的所有差异。使用线性模型是基于以下两点考虑：1）其他的工作中，线性模型被验证是有效的，如在标记分布学习中假设线性模型作为特征空间到标记空间的映射的工作取得了良好的预测效果；2）线性模型具有运算速度快和便于求解的特点。映射到标记空间的两种分布的经验均值之间的差异表示为：

$$\left\|\frac{1}{n_s}\sum_{i\in S}f(\boldsymbol{x}_i)-\frac{1}{n_p}\sum_{j\in P}f(\boldsymbol{x}_j)\right\|_F^2$$

即最大均值差异，经证明它是有效衡量其边际概率分布差异的方法。

文中用最大均值差异来衡量训练集与未标注数据集示例之间的分布差异，选择 b 个示例的数据集 Q 进行查询，使所选示例代表的边际分布最接近未标记数据所代表的分布。通过使训练数据集上的损失最小化所获得的模型对未标记数据以及来自同一分布的数据具有良好的泛化能力。引入集合 $L\cup Q$ 和 $U\backslash Q$ 之间的 MMD，式(1)改为：

$$\sum_{i\in L}\|f(\boldsymbol{x}_i)-\boldsymbol{y}_i\|_F^2+\lambda\left\|\frac{1}{n_l+b}\sum_{i\in L\cup Q}f(\boldsymbol{x}_i)-\frac{1}{n_u-b}\sum_{i\in U\backslash Q}f(\boldsymbol{x}_j)\right\|_F^2 \quad (2)$$

由于我们要选择的是使 $L\cup Q$ 分布与 $U\backslash Q$ 分布差别最小化的 Q，因此引入了一个稀疏长度为 n_u 的向量 $\boldsymbol{\alpha}$，每一项 α_j 为 0 或 1。如果在未标注数据集 U 中选择了一个示例，则这个示例对应的元素 α_j 为 1，否则为 0。以上最小化问题即转化为优化向量 $\boldsymbol{\alpha}$ 的问题：

$$\min_{\boldsymbol{\alpha}}\sum_{i\in L}\|f(\boldsymbol{x}_i)-\boldsymbol{y}_i\|_F^2+\lambda\left\|\frac{\sum_{i\in L}f(\boldsymbol{x}_i)+\sum_{j\in U}\alpha_j f(\boldsymbol{x}_i)}{n_l+b}-\frac{\sum_{i\in U}(1-\alpha_j)f(\boldsymbol{x}_j)}{n_u-b}\right\|_F^2 \quad (3)$$

$$\text{s.t. } \alpha_j\in\{0,1\},\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=b$$

由于我们假设从特征空间到标记空间的映射是 $f(\boldsymbol{X})=\boldsymbol{WX}$，代入上式可得：

$$\min_{\boldsymbol{W},\boldsymbol{\alpha}}\sum_{i\in L}\|(\boldsymbol{W}\boldsymbol{x}_i-\boldsymbol{y}_i)\|_F^2+\lambda\left\|\frac{1}{n_l+b}\left(\sum_{i\in L}\boldsymbol{W}\boldsymbol{x}_i+\sum_{j\in U}\alpha_j\boldsymbol{W}\boldsymbol{x}_j\right)-\frac{1}{n_u-b}\sum_{i\in U}(1-\alpha_j)\boldsymbol{W}\boldsymbol{x}_j\right\|_F^2$$

$$\text{s.t. } \alpha_j\in\{0,1\},\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=b \quad (4)$$

其中，1 是与 $\boldsymbol{\alpha}$ 具有相同维数且所有元素均为 1 的向量。第一项表示训练一个线性回归模型，使模型的训练损失最小；第二项表示加上所选查询集的新训练集的预测标记的平均值与所选查询集的新未标记数据集的预测标记的平均值之差。第一个约束条件确保向量 $\boldsymbol{\alpha}$ 中的每个元素均为 0 或 1；第二个约束条件确保向量 $\boldsymbol{\alpha}$ 中正好有 b 个元素为 1，表示从未标记的数据集中选择了 b 个示例。

2.3 求解

优化问题(4)的求解需要找到使目标最小化的两个变量：矩阵 $\boldsymbol{W}$ 和向量 $\boldsymbol{\alpha}$。可以采用交替迭代的方式求解。

1)优化矩阵 $\boldsymbol{W}$

固定向量 $\boldsymbol{\alpha}$，则目标可以改写为以下形式：

$$\min_{\boldsymbol{W}} \| \boldsymbol{WX}-\boldsymbol{Y} \|_F^2+\lambda \| \boldsymbol{WM} \|_F^2 \tag{5}$$

其中，$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_{n_l}]$，$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_c]$，$\boldsymbol{M}=\frac{1}{n_l+b}\sum_i \boldsymbol{x}_i+\frac{1}{n_l+b}\sum_j \alpha_j \boldsymbol{x}_j-\frac{1}{n_u-b}\sum_j (1-\alpha_j)\boldsymbol{x}_j$。这是一个有闭式解的凸问题，可以通过求极值来求解。

对目标函数求导，在驻点目标函数的导数为 0，即：

$$(\boldsymbol{WX}-\boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}+\lambda \boldsymbol{WMM}^{\mathrm{T}}=0 \tag{6}$$

进一步解得：

$$\boldsymbol{W}=\boldsymbol{YX}^{\mathrm{T}}(\boldsymbol{XX}^{\mathrm{T}}+\lambda \boldsymbol{MM}^{\mathrm{T}})^{-1} \tag{7}$$

此时，驻点即为该凸函数中的最小值点。

2)优化向量 $\boldsymbol{\alpha}$

固定矩阵 $\boldsymbol{W}$，则优化目标可以改写为：

$$\min_{\boldsymbol{\alpha}} \left\| \frac{1}{n_l+b}\left(\sum_{i\in L}\boldsymbol{Wx}_i+\sum_{j\in U}\alpha_j \boldsymbol{Wx}_j\right)-\frac{1}{n_u-b}\sum_{i\in U}(1-\alpha_j)\boldsymbol{Wx}_j \right\|_F^2$$
$$\text{s.t. } \alpha_j\in\{0,1\}, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b \tag{8}$$

这是个二次公式，可以重新表述为整数线性规划问题。将优化目标(8)进一步拆分为：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{\alpha}-\boldsymbol{k}_2^{\mathrm{T}}\boldsymbol{\alpha}+\boldsymbol{k}_3^{\mathrm{T}}\boldsymbol{\alpha}+const$$
$$\text{s.t. } \alpha_j\in\{0,1\}, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}= b \tag{9}$$

令矩阵 $\boldsymbol{G}$ 为定义在未标注数据集 U 和训练集 L 上的 $(n_u+n_l)\times(n_u+n_l)$ Gram 矩阵，且 $\boldsymbol{G}(i,j)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x}_j$。定义矩阵 $\boldsymbol{K}_1=\boldsymbol{G}(1:n_u,1:n_u)$，向量 $\boldsymbol{k}_2(i)=\frac{n_l+b}{n_l+n_u}\sum_{i=1}^{n_u}\boldsymbol{K}_1(i,j)$ 和向量 $\boldsymbol{k}_3(i)=\frac{n_u-b}{n_l+n_u}\sum_{i=1}^{n_l}\boldsymbol{G}(i,n_u+j)$。

化简后的优化问题(9)的第一项确保了查询集在其内部具有最小的相似性，从而避免了选择相似点造成的冗余；第二项表示选中的示例与未选中的示例相似，从而确保选择了有代表性的点；第三项表示与已标记数据的相似度较低的示例更有可能被选中，使训练集中的数据具有多样性。

由于 $\alpha_j\in\{0,1\}$，以上目标函数线性项可以与二次项矩阵合并成矩阵 $\boldsymbol{D}$，则 $i=j$ 时，$\boldsymbol{D}_{ji}=\boldsymbol{K}_1(j,i)-\boldsymbol{k}_2(j)+\boldsymbol{k}_3(j)$；$i\neq j$ 时，$\boldsymbol{D}_{ji}=\boldsymbol{K}_1(j,i)$。因此，目标函数可以改写为：

$$\min_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\alpha}$$
$$\text{s.t. } \alpha_j\in\{0,1\}, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b \tag{10}$$

引入矩阵 $\boldsymbol{Z}_{j\times i}$，其元素满足 $z_{ji}=\alpha_j\alpha_i$，则优化问题(8)可以改写为：

$$\min_{\boldsymbol{\alpha},\boldsymbol{Z}} \sum_{i,j} d_{ji}z_{ji}$$
$$\text{s.t. } z_{ji}=\alpha_j\alpha_i, \alpha_j\in\{0,1\}, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b \tag{11}$$

由于 d_{ij} 可以是正值，也可以是负值，重写约束如下：

$$\min_{\boldsymbol{\alpha},\boldsymbol{Z}} \sum_{i,j} d_{ji}z_{ji}$$
$$\text{s.t. } -\alpha_i-\alpha_j+2z_{ji}\leqslant 0 \text{ for } d_{ji}<0$$
$$\alpha_i+\alpha_j-z_{ji}\leqslant 1 \text{ for } d_{ji}\geqslant 0 \tag{12}$$
$$\alpha_j\in\{0,1\}, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b$$

如果 α 或 α_j 等于零，则第一项约束确保了 $z_{ji}=0$。如果 α_i 或 α_j 都等于 1，由于 d_{ij} 取负值，则 $z_{ji}=1$。当 d_{ij} 为正时，如果 α_i 或 α_j 都等于 1，则第二个约束条件保证了 $z_{ji}=1$；如果 α_i 或 α_j 等于零，则 z_{ji} 等于 0。因此，当且仅当 α_i 或 α_j 都等于 1 时，$z_{ji}=1$。最终，优化目标(12)把优化目标(11)转化为了一个整数线性规划问题。

下面提出两种方法，分别解决优化问题(9)和优化问题(12)中定义的整数二次规划和整数线性优化问题。具体方式是通过松弛整数约束，把优化问题(9)和优化问题(12)分别放缩为二次规划(QP)和线性规划(LP)问题。

由于 $\alpha_j\in\{0,1\}$ 是二进制约束，因此式(9)是一个整数二次规划问题，更是一个 NP-hard 问题。通过采取松弛约束的方法，使其成为一个连续的二次规划问题，从而得出以下公式：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{\alpha}-\boldsymbol{k}_2^{\mathrm{T}}\boldsymbol{\alpha}+\boldsymbol{k}_3^{\mathrm{T}}\boldsymbol{\alpha}+const$$
$$\text{s.t. } \alpha_j\in[0,1], \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b \tag{13}$$

同理，通过将第一项约束条件纳入目标函数，可以进一步简化式(13)。当 d_{ji} 取负值时，由第一等式约束 $z_{ji}=\frac{\alpha_j+\alpha_i}{2}$。但是，当 d_{ji} 取 0 或正值时，第二个等式约束可能无法保持。由于移除或放松最小化程序的约束不会增加最优值，式(12)可以重新表示为以下连续的线性规划问题。

$$\min_{\boldsymbol{\alpha},\boldsymbol{Z}} \frac{1}{2}\sum_{d_{ji}<0} d_{ji}(\alpha_i+\alpha_j)+\sum_{d_{ji}\geqslant 0} d_{ji}z_{ji}$$
$$\text{s.t. } \alpha_i+\alpha_j-z_{ji}\leqslant 1 \text{ for } d_{ji}\geqslant 0$$
$$z_{ji}\in[0,1] \text{ for } d_{ji}\geqslant 0 \tag{14}$$
$$\alpha_j\in[0,1], \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=b$$

优化问题(14)的线性规划问题等效于优化目标(13)的二次规划问题。由于优化问题(9)中的 $\boldsymbol{K}_1$ 是正半定数的核 Gram 矩阵，因此两个公式都是凸的。可以使用标准的 LP 解算器 CVX[21] 来解决 LP 问题，或者使用 Matlab 自带的 quadprog 函数来解决 QP 问题。本文使用 quadprog 函数来解决 QP 问题，以求解向量 $\boldsymbol{\alpha}$。如算法 1 所示，在一轮迭代选点中，在未标记集 U 中根据更新的向量 $\boldsymbol{\alpha}$ 选择 b 个示例，查询其标签投入训练集 L，从而使训练集 L 与未标记集 U 都得到更新。

算法 1

输入：训练集 L，未标注数据集 U，选点个数 b，参数 λ；

输出：数据集 Q

1. 初始化：向量 $\alpha_j=\frac{1}{n_u}$

反复迭代以下步骤：

2. 根据式(7)更新矩阵 W；

3. 根据式(13)更新向量 $\boldsymbol{\alpha}$，直至满足收敛条件结束；

4. 以 $\boldsymbol{\alpha}$ 的降序对 U 进行排序，然后选择前 b 个示例作为 Q(即选择 $\boldsymbol{\alpha}$

最大的 b 个元素所对应的 U 中的示例作为 Q)；

5. 更新集 L 和 U：L→L∪Q，U→U\Q

3 理论分析

3.1 收敛性分析

性质 1 经过算法 1 优化，优化问题(4)中的目标函数值在每次迭代中不会增加。

证明：假设目标函数经过 t 轮，迭代为 $obj(\boldsymbol{W}^{(t)},\boldsymbol{\alpha}^{(t)})$。

当固定向量 $\boldsymbol{\alpha}$ 时，优化矩阵 $\boldsymbol{W}$，求解优化问题(5)。由于优化问题(5)是凸的，求得 $\boldsymbol{W}^{(t+1)}$ 是使优化问题(5)最小的最优解，因此更新矩阵 $\boldsymbol{W}$ 不会增加目标函数值，即 $\text{obj}(\boldsymbol{W}^{(t+1)},\boldsymbol{\alpha}^{(t)})\leqslant obj(\boldsymbol{W}^{(t)},\boldsymbol{\alpha}^{(t)})$。

当固定矩阵 $\boldsymbol{W}$ 时，优化向量 $\boldsymbol{\alpha}$，求解最小化问题式(8)。由于优化问题(8)的子问题式(13)是凸的，求得 $\boldsymbol{\alpha}^{(t+1)}$ 是使式(13)最小的最优解，因此更新向量 $\boldsymbol{\alpha}$ 不会增加式(13)的目标函数值，即 $obj(\boldsymbol{W}^{(t+1)},\boldsymbol{\alpha}^{(t+1)})\leqslant obj(\boldsymbol{W}^{(t+1)},\boldsymbol{\alpha}^{(t)})$。

由于移除或放松最小化式的约束不会增加最优值，因此在放缩前函数同样满足 $obj(\boldsymbol{W}^{(t+1)},\boldsymbol{\alpha}^{(t+1)})\leqslant obj(\boldsymbol{W}^{(t)},\boldsymbol{\alpha}^{(t)})$。

因此，目标函数值在迭代优化过程中不会增加。

3.2 时间复杂度分析

可以采用计算每个子步骤的时间复杂度并叠加的方法计算 ALDL-MMD 模型每一轮选点的时间复杂度。每个子步骤的时间复杂度如下：

(1)固定向量 $\boldsymbol{\alpha}$，优化矩阵 $\boldsymbol{W}$，求解优化问题(5)，其时间复杂度为 $O(n_u^2d+cn_ld^4)$；

(2)固定矩阵 $\boldsymbol{W}$，优化向量 $\boldsymbol{\alpha}$，求解优化问题(8)，其时间复杂度为 $O((n_l+2n_u)^2)$；

(3)以 $\boldsymbol{\alpha}$ 的降序对 U 进进行排序，其时间复杂度为 $O(n_u)$。

假设目标函数经过 t 轮迭代收敛，则 ALDL-MMD 总的时间复杂度为 $O(t\times(n_u^2d+cn_ld^4+(n_l+2n_u)^2))$。

4 实验设计与结果分析

4.1 数据集概述

实验中总共使用了 7 个数据集。表 2 列出了这些数据集的一些基本统计信息，包括示例数、标记数和特征数。

表 2 数据集

Table 2 Datasets

数据集	示例数	特征数	标记数
Yeast-cold	2 465	24	4
Yeast-spo5	2 465	24	3
Yeast-diau	2 465	24	7
Yeast-heat	2 465	24	6
Yeast-dtt	2 465	24	4
Yeast-spoem	2 465	24	2
Yeast-spo	2 465	24	6

这 7 个数据集是从酿酒酵母 Saccharomyces cerevisiae 的生物学实验中收集的真实数据集[22]。每个数据集记录一个实验的结果，总共包括 2465 个酵母菌基因，每个基因均由长度为 24 的相关系统发育谱载体表示。对于每个数据集，标记表示一个生物学实验中的离散时间点。在每个时间点的基因表达水平给出了相应标记的描述程度的自然度量。基因表达水平的值是标准化后的，提供了相应标记描述程度的自然度量。所有标记(时间点)的描述程度(标准化基因表达水平)构成了特定酵母基因的标记分布。

4.2 实验方案

将所提算法与已有的主动学习示例选择方法进行对比，并用标记分布学习的 SA-BFGS 算法进行预测。实验比较了以下 5 种主动学习方法在标记分布数据集中的结果。

(1)Random：随机选择示例查询其标记的基线方法。

(2)Maximin-based Anomaly Detection(MMAD)[22]：将版本空间近似为涵盖大多数假设的结构化超球面，然后将可用的采样方法转换为内部体积采样，是一种无监督的主动选点算法。

(3)Manifold Adaptive Experimental Design(MAED)[17]：用于数据流形自适应内核空间。在流形自适应内核空间中，数据的流形结构以图拉普拉斯算子形式存在于内核空间中。通过最大程度地减小针对最佳分类器的预测损失，可以选择最具代表性和区分性的示例进行标记，是一种无监督的主动选点算法。

(4)Query-By-Committee (QBC)[7]：采用最大预测不确定性作为主动选择示例的标准。所构建的委员会成员分别是 PT-Bayes，PT-SVM，AA-kNN，AA-BP，LD-SVR，CPNN，SA-IIS 和 SA-BFGS[2] 算法。QBC 算法是一种监督的主动选点算法。

(5)ALDL-MMD：本文提出的方法。

我们将每个数据集随机分为 3 部分：第一部分是初始标记的训练集；第二部分是带有标记的示例的测试集；第三部分是用于示例选择的未标记示例集。在开始进行主动学习实验时，随机选择 50 个示例作为初始标记训练数据。在主动学习的每次迭代中，主动学习方法的每轮选点会根据各自的策略选择 10 个示例进行查询(即每轮选点的查询集的示例数为 10)，然后将其添加到训练集中，在标记的数据(训练集)上训练主动学习模型，并使用 2 个测量值(Canberra Metric 和 Intersection[23])在测试数据上评估其性能。将所有的未标记数据添加到训练集中后，查询过程将停止。以上过程重复 10 次，取 10 次重复实验的平均结果作最终结果。

选择 SA-BFGS 算法对预测数据进行预测。输出的预测数据的标记是标记分布，评估标记分布标注是否准确的方法是衡量预测标记和真实标记分布之间的相似性或距离。对于概率分布之间的距离/相似度，有一些度量标准可以很好地用于测量标记分布之间的距离/相似度。本文评估了 5 种比较方法在“Canberra Metric”和“Intersection”度量指标上的效果。

假设预测标记分布和真实标记分布分别为 $\hat{\boldsymbol{D}}$ 和 $\boldsymbol{D}$，$d_x^{y_j}$ 是第 j 个标签对示例的描述度。“Canberra Metric”是一个概率分布之间的距离指标，表达式为 $Canberra(\boldsymbol{D},\hat{\boldsymbol{D}})=\sum_{j=1}^{c}\frac{|d_x^{y_j}-\hat{d}_x^{y_j}|}{d_x^{y_j}+\hat{d}_x^{y_j}}$；“Intersection”是一个相似度指标，表达式为 $Intersection(\boldsymbol{D},\hat{\boldsymbol{D}})=\min(d_x^{y_j},\hat{d}_x^{y_j})$。

4.3 实验结果分析

图 1 所示为 ALDL-MMD 和对比方法在 Canberra Metric 指标上的结果。查询示例的 5 种方法每轮选点的预测数据的预测标记分布与预测数据的真实标记分布的 Canberra Metric 值用不同颜色和形状的点表示,并用不同颜色的虚线连接。

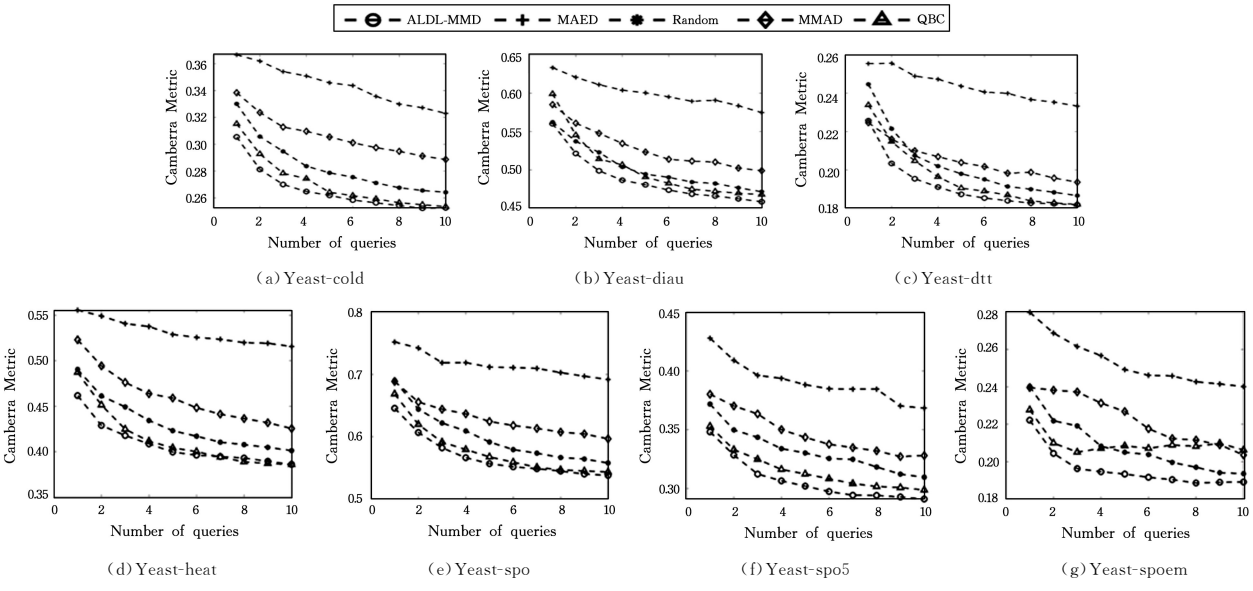


图 1 真实数据集上由 Canberra Metric ↓ 衡量的实验结果("↓"表示指标越小越好)(电子版为彩色)

Fig. 1 Experimental results on real-world datasets measured by Canberra Metric ↓ (↓ means the small the better)

可以看出,不同算法选择的示例对模型效果提升的贡献不同。总体而言,在 7 个数据集上,ALDL-MMD 方法比其他对比方法更有效。此外,在所有数据集中,随机并不是最差的方法。造成此现象的原因可能是:随机方法查询示例及其标记,而 MMAD 和 MAED 是两种无监督的主动选点方法。标记分布学习的核心在于建立特征空间到标记空间的有效映射,而且关联程度高于一般的分类学习。特征空间的数据结构与标记之间的关联更高,从而无法从分析特征空间数据结构的角度来有效选点。有监督的主动选点方法比无监督的主动选点方法表现出更好的性能。另一个原因可能与标记相关,同一示例的标记之间可能存在信息重复。这就解释了为什么在某些数据上随机选择会比一些主动方法更好。作为查询示例及其标记的方法,ALDL-MMD 和 QBC 往往比只查询示例的 MAED 和 MMAD 方法更有效。

当查询次数小于 7 时,不同方法间的效果差距较为明显。当查询的次数增加到 10 时,所有模型通常都会变得更好。通过不同的主动学习方法选择的 100 个点对改善模型的性能做出了不同程度的贡献。这同样也说明了采用不同的算法效果也会不同。当训练集的示例达到一定数目时,再增加训练数据,模型效果的提升不再明显,此时本文算法效果会接近于对比方法的效果。

表 3 列举了查询次数分别为 1,3,5,7,9 时,7 个数据集上不同主动学习算法的 Intersection 指标的性能,其中展示的是 10 次实验的平均效果和标准差。对于每种情况,根据 95%显著性水平成对 t 检验,将最佳对比结果加粗。

表 3 真实数据集上 Intersection ↑ 指标衡量的实验结果(均值±标准差)("↑"表示指标越大越好)

Table 3 Experimental results (mean±std) on real-world datasets measured by Intersectio ↑ (↑ means the bigger the better)

数据集	算法	查询次数				
		1	3	5	7	9
Yeast-cold	Random	0.9183±0.0016	0.9271±0.0017	0.9311±0.0031	0.9330±0.0033	0.9344±0.0017
	MAED	0.9095±0.0040	0.9126±0.0025	0.9146±0.0032	0.9171±0.0022	0.9192±0.0039
	MMAD	0.9164±0.0011	0.9227±0.0023	0.9245±0.0010	0.9264±0.0022	0.9279±0.0029
	QBC	0.9220±0.0022	0.9312±0.0029	**0.9348±0.0056**	0.9360±0.0034	**0.9371±0.0031**
	ALDL-MMD	**0.9243±0.0020**	**0.9333±0.0015**	**0.9353±0.0043**	**0.9368±0.0006**	**0.9378±0.0026**
Yeast-spo	Random	0.6901±0.0022	0.6220±0.0026	0.5915±0.0010	0.5736±0.0016	0.5642±0.0020
	MAED	0.75196±0.0029	0.7186±0.0010	0.7118±0.0034	0.7097±0.0040	0.6970±0.0015
	MMAD	0.6886±0.0029	0.6441±0.0011	0.6245±0.0022	0.6133±0.0023	0.6044±0.0031
	QBC	0.6686±0.0043	**0.5913±0.0036**	**0.5676±0.0024**	**0.5509±0.0035**	**0.5450±0.0026**
	ALDL-MMD	**0.6456±0.0016**	**0.5822±0.0033**	**0.5568±0.0022**	**0.5487±0.0022**	**0.5402±0.0017**
Yeast-dtt	Random	**0.9395±0.0016**	0.9487±0.0038	0.9510±0.0042	**0.9527±0.0022**	**0.9535±0.0020**
	MAED	0.9367±0.0014	0.9382±0.0017	0.9395±0.0039	0.9404±0.0026	0.9416±0.0015
	MMAD	**0.9446±0.0020**	0.9482±0.0035	0.9497±0.0015	0.9511±0.0010	0.9517±0.0011
	QBC	0.9421±0.0040	**0.9494±0.0007**	0.9529±0.0006	0.9538±0.0040	0.9548±0.0029
	ALDL-MMD	**0.9442±0.0031**	**0.9518±0.0039**	**0.9538±0.0039**	**0.9546±0.0031**	**0.9550±0.0015**

（续表）

数据集	算法	查询次数				
		1	3	5	7	9
Yeast-spo5	Random	0.8795±0.0016	0.8886±0.0015	0.8931±0.0020	0.8948±0.0007	0.8988±0.0040
	MAED	0.8627±0.0022	0.8722±0.0036	0.8748±0.0043	0.8761±0.0031	0.8805±0.0011
	MMAD	0.8768±0.0016	0.8822±0.0017	0.8886±0.0037	0.8913±0.0038	0.8939±0.0039
	QBC	0.8854±0.0015	0.8946±0.0029	**0.8987±0.0032**	0.9014±0.0027	0.9025±0.0038
	ALDL-MMD	**0.8870±0.0036**	**0.8986±0.0039**	**0.9020±0.0011**	**0.9046±0.0034**	**0.9051±0.0040**
Yeast-spoem	Random	0.8847±0.0035	0.8944±0.0010	0.9009±0.0016	**0.9035±0.0040**	**0.9061±0.0047**
	MAED	0.8666±0.0020	0.8746±0.0018	0.8802±0.0043	0.8818±0.0036	0.8838±0.0006
	MMAD	0.8844±0.0024	0.8856±0.0039	0.8904±0.0033	0.8972±0.0026	0.8988±0.0015
	QBC	0.8901±0.0028	**0.9007±0.0038**	0.8991±0.0038	0.8988±0.0015	0.8983±0.0040
	ALDL-MMD	**0.8925±0.0020**	**0.9050±0.0010**	**0.9064±0.0043**	**0.9079±0.0029**	**0.9086±0.0026**
Yeast-diau	Random	**0.9214±0.0043**	0.9270±0.0034	0.9311±0.0042	0.9325±0.0044	0.9336±0.0038
	MAED	0.9115±0.0029	0.9143±0.0030	0.9159±0.0020	0.9175±0.0039	0.9185±0.0017
	MMAD	0.9182±0.0022	0.9235±0.0028	0.9270±0.0020	0.9287±0.0020	0.9299±0.0030
	QBC	0.9163±0.0015	**0.9285±0.0017**	**0.9319±0.0014**	**0.9341±0.0012**	0.9348±0.0045
	ALDL-MMD	**0.9217±0.0026**	**0.9305±0.0012**	**0.9331±0.0017**	**0.9349±0.0027**	**0.9358±0.0022**
Yeast-heat	Random	0.9189±0.0010	0.9259±0.0023	0.9303±0.0039	0.9324±0.0044	0.9334±0.0031
	MAED	0.9082±0.0038	0.9106±0.0039	0.9126±0.0015	0.9134±0.0017	0.9141±0.0015
	MMAD	0.9140±0.0015	0.9216±0.0040	0.9245±0.0022	0.9274±0.0020	0.9290±0.0040
	QBC	0.9194±0.0010	0.9300±0.0012	0.9334±0.0029	**0.9352±0.0031**	**0.9363±0.0029**
	ALDL-MMD	**0.9238±0.0022**	**0.9313±0.0031**	**0.9343±0.0016**	**0.9352±0.0041**	**0.9361±0.0041**

表 4 显示了查询次数不同时，ALDL-MMD 相比于其他方法的 Intersection 指标的胜/平/输的次数。

表 4 不同查询次数下 ALDL-MMD 相比于对比方法的胜/平/输次数

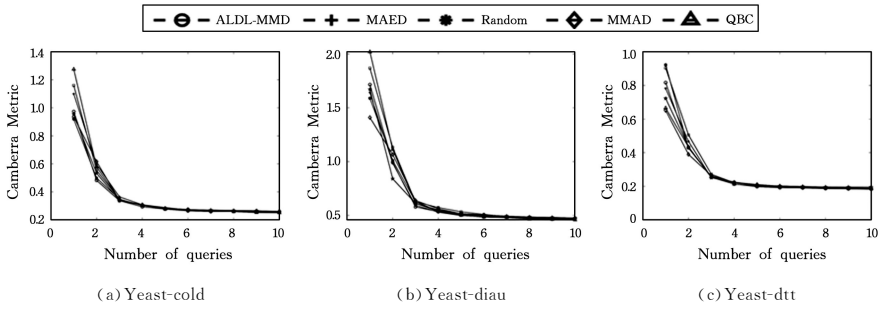
Table 4 Win/tie/loss counts of ALDL-MMD versus comparative methods with varied numbers of queries

算法	查询次数					总计
	1	3	5	7	9	
Random	5/2/0	7/0/0	7/0/0	5/2/0	5/2/0	29/6/0
MAED	7/0/0	7/0/0	7/0/0	7/0/0	7/0/0	35/0/0
MMAD	6/1/0	7/0/0	7/0/0	7/0/0	7/0/0	34/1/0
QBC	7/0/0	3/4/0	3/4/0	4/3/0	3/4/0	20/15/0
总计	25/3/0	24/4/0	24/4/0	23/5/0	22/6/0	118/22/0

可以看出，在大多数情况下，ALDL-MMD 优于其他 4 种对比方法。在所有数据集上，ALDL-MMD 的表现都不比其他对比方法差，说明了 ALDL-MMD 的有效性。当查询次数大于 7 时，QBC 和 MMAD 方法的表现逐渐接近 ALDL-MMD 的表现，这同样说明了在标记分布学习中，训练集的数据数明显增加并不代表着标记分布学习的性能也会明显提高。5 种对比方法显示了主动学习研究的两个方向：选择信息最丰富的示例和选择最具代表性的示例。MAED，MMAD 和 ALDL-MMD 选择由先前训练的模型给出的最具代表性的示例，而 QBC 选择信息最丰富的示例。在本实验中，对比方法效果的好坏与是哪种方向的方法没有必然联系。

4.4 参数分析

图 2 分别展示了在 Yeast-cold，Yeast-diau 和 Yeast-dtt 数据集上参数 λ 对 ALDL-MMD 选点方法效果 Canberra Metric 指标的影响。采用网格法选取参数，参数 λ 的取值设定为 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$。可以看出，在 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$ 这个较大的跨度内，随着参数的变化，主动选点方法的效果变化并不明显，说明 ALDL-MMD 比较稳定，模型精度受参数的影响较小，这符合我们对模型的最初设定。参数的取值对主动选点实验结果影响不大，说明 ALDL-MMD 方法对参数选择的要求不高，可用性更高。



(a) Yeast-cold (b) Yeast-diau (c) Yeast-dtt

图 2 真实数据集上不同参数由 Canberra Metric ↓ 衡量的实验结果

Fig. 2 Experimental results on the real-world datasets measured by Canberra Metric ↓

结束语 本文旨在解决标记分布学习中所需训练数据的高额标注成本问题，提出了一种新的算法 ALDL-MMD。该方法训练一个线性回归模型，使该模型的训练损失最小，同时优化一个旨在选择示例的稀疏的二进制向量，使得每次选点中的训练数据与无标记数据的数据分布尽可能相似，并且选择查询的数据最有代表性，从而提升模型的效果。文中还进一步分析了算法的收敛性和时间复杂度。在真实数据集上的实验结果表明，ALDL-MMD 算法比对比方法更有效。另外，还通过实验展示了参数选择对实验结果的影响。

接下来的工作是进一步在主动标记分布学习中建立示例信息量和代表性的准则，这需要我们对标记分布数据的数据结构有进一步的认识，并对特征到标记映射的其他理论进行

进一步研究。我们未来的另一项任务是系统地研究标记分布最大熵模型理论,从而设计出更适用于最大熵模型的主动标记分布示例选择算法,以进一步提升主动标记分布算法的性能。

参 考 文 献

[1] ZHANG M,ZHOU Z. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge & Data Engineering,2014,26(8):1819-1837.

[2] GENG X,JI R. Label Distribution Learning[C]// IEEE International Conference on Data Mining Workshops. IEEE Computer Society,2013.

[3] HE Z,LI X,ZHANG Z. Data-Dependent Label Distribution Learning for Age Estimation[J]. IEEE Trans. Image Process.,2017,26(8):3846-3858.

[4] KONG S,OYINI MBOUNA R. Head Pose Estima-tion from a 2-D Face Image using 3-D Face Morphing with Depth Parameters [J]. IEEE Transactions on Image Processing,2015,24(6):1-1.

[5] ZHANG Z L,LAI C H,LIU H,et al. Infrared Facial Expression Recognition via Gaussian-based Label Distribution Learning in the Dark Illumination Environment[J]. Neurocomputing,2020,409:341-350.

[6] ZHOU D Y,ZHANG X,ZHOU Y,et al. Emotion Distribution Learning from Texts[C]// Conference on Empirical Methods in Natural Language Processing. 2016.

[7] SEUNG H S. Query by Committee[C]// Workshop on Computational Learning Theory. ACM,1992.

[8] FREUND Y,SEUNG H S,SHAMIR E,et al. Selective Sampling Using the Query by Committee Algorithm[J]. Machine Learning,1997,28(2/3):133-168.

[9] GU S,CAI Y,SHAN J,et al. Active Learning with Error-Correcting Output Codes[J]. Neurocomputing,2019,364:182-191.

[10] LINDLEY D V. On a Measure of the Information Provided by an Experiment [J]. Annals of Mathematical Statistics,1956,27(4):986-1005.

[11] YU K,BI J B,TRESP V,et al. Active learning via transductive experimental design[C]// International Conference on Machine Learning. 2006:1081-1088.

[12] PATRA S,BRUZZONE L. A cluster-assumption based batch mode active learning technique[J]. Pattern Recognition Letters,2012,33(9):1042-1048.

[13] BURGES C J. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery,1998,2(2):121-167.

[14] MCCALLUM A,NIGAM K. A comparison of event models for naive bayes text classification[C]// National Conference on Artificial Intelligence. 1998:41-48.

[15] PIETRA S D,PIETRA V D,LAFFERTY J,et al. Inducing features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1997,19(4):380-393.

[16] NOCEDAL J,WRIGHT S. Numerical optimization[M]. Springer Science & Business Media,2006:61-63.

[17] PRINCE M. Does Active Learning Work? A Review of the Research[J]. Journal of Engineering Education,2004,93(3):223-231.

[18] CAI D,HE X. Manifold Adaptive Experimental Design for Text Categorization[J]. IEEE Transactions on Knowledge and Data Engineering,2011,24(4):707-719.

[19] SINDHWANI V,NIYOGI P,BELKIN M,et al. Beyond the point cloud:from transductive to semi-supervised learning[C]// International Conference on Machine Learning. 2005:824-831.

[20] PAN S J,TSANG I W,KWOK J T,et al. Domain Adaptation via Transfer Component Analysis [J]. IEEE Transactions on Neural Networks,2011,22(2):199-210.

[21] GRANT M,BOYD S. CVX:Matlab software for disciplined convex programming[J]. International Journal of Communications,Network and System Science,2008,1(1).

[22] EISEN M B,SPELLMAN P T,BROWN P O,et al. Cluster analysis and display of genome-wide expression patterns[J]. Proceedings of the National Academy of Sciences of the United States of America,1998,95(25):14863-14868.

[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.

DONG Xin-yue, born in 1996, postgraduate. Her main research interests include statistical data analysis and machine learning.

HOU Chen-ping, born in 1982, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning, statistical data analysis, pattern recognition and computer vision.