# 最大规范化依赖性多标记半监督学习方法

张晨光 [1] 张 燕 [1] 张夏欢 [2]

**摘 要** 针对现有多标记学习方法大多属于有监督学习方法, 而不能有效利用相对便宜且容易获得的大量未标记样本的问题, 本文提出了一种新的多标记半监督学习方法, 称为最大规范化依赖性多标记半监督学习方法 (Normalized dependence maximization multi-label semi-supervised learning method). 该方法将已有标签作为约束条件, 利用所有样本, 包括已标记和未标记样本, 对特征集和标签集的规范化依赖性进行估计, 并以该估计值的最大化为目标, 最终通过求解带边界的迹比值问题为未标记样本打上标签. 与其他经典多标记学习方法在多个真实多标记数据集上的对比实验表明, 本文方法可以有效从已标记和未标记样本中学习, 尤其是已标记样本相对稀少时, 学习效果得到了显著提高.

**关键词** 规范化依赖性, 多标记学习, 半监督学习, 迹比值

**引用格式** 张晨光, 张燕, 张夏欢. 最大规范化依赖性多标记半监督学习方法. 自动化学报, 2015, **41**(9): 1577−1588

**DOI** 10.16383/j.aas.2015.c140893

## Normalized Dependence Maximization Multi-label Semi-supervised Learning Method

ZHANG Chen-Guang[1] ZHANG Yan[1] ZHANG Xia-Huan[2]

**Abstract** In view of the problems that most of present multi-label learning methods are supervised learning methods and cannot effectively make use of relatively inexpensive and easily obtained large number of unlabeled samples, this paper puts forward a new multi-label semi-supervised learning method, called normalized dependence maximization multi-label semi-supervised learning method (DMMS). The DMMS regards labeled samples as constraint conditions, estimates the normalized dependency of feature and label sets on all samples including labeled and unlabeled samples, and maximizes the estimation by finally addressing a trace ratio optimization problem with constraint conditions for label unlabeled samples. Experiments comparing DMMS with the state-of-the-art multi-label learning approaches on several real-world datasets show that the DMMS can effectively learn from labeled and unlabeled samples, especially when the labeled is relatively rare, the learning performance can be improved greatly.

**Key words** Normalized dependence, multi-label learning, semi-supervised learning, trace ratio

**Citation** Zhang Chen-Guang, Zhang Yan, Zhang Xia-Huan. Normalized dependence maximization multi-label semi-supervised learning method. *Acta Automatica Sinica*, 2015, **41**(9): 1577−1588

传统的分类学习, 包括多类学习 (Multi-class learning) 每个样本只属于一个类别, 可统称为单标记学习问题 (Single-label learning). 然而, 实际应用中, 一个样本可能同时属于多个类别. 例如, 一篇文档可能属于多个预定义的主题; 一张图片可能同时具有多个语义; 一个基因可能具有多种功能. 这种单个样本具有多个类别的学习问题称为多标记学习问题. 传统的学习方法, 包括近邻法、决策树、神经网络和支持向量机等都不能直接用于多标记学习问题. 为此, 研究者提出了问题转换和算法改进两种解决方案[1−2]. 问题转换法主要针对标签集, 通过处理样本标签将多标记学习问题转换成传统学习方法能解决的问题. 比如, Binary relevance[2] 和 Classifier chains[3] 对每个标签都分别建立分类器, 将多标记学习问题转换成若干二分类问题; Label powerset[4] 和 Random $k$-labelsets[5] 将样本标签组合视为新类标签, 把多标记问题转换成多类学习问题. 算法改进方法则是通过改进单标记学习方法使之适用于多标记学习问题. 比如, Multi-label $k$-nearest neighbor (MLKNN)[6] 和 Rank-SVM[7] 分别改进了 $k$ 近邻和支持向量机, 使之适用于多标记学习.

目前, 无论是基于问题转换还是算法改进的多标记学习方法, 研究重点多集中在有监督学习范畴. 一般地, 有监督学习方法往往需要足量的已标

收稿日期 2015-01-13 录用日期 2015-05-06
Manuscript received January 13, 2015; accepted May 6, 2015
国家自然科学基金 (11261015), 海南省高等学校科学研究项目 (Hjkj2012-01) 资助
Supported by National Natural Science Foundation of China (11261015), College Scientific Research Program of Hainan Province of China (Hjkj2012-01)
本文责任编委 周志华
Recommended by Associate Editor ZHOU Zhi-Hua
1. 海南大学信息科学技术学院 海口 570228 2. 北京凌云光视公司图像处理部 北京 100097
1. College of Information Science and Technology, Hainan University, Haikou 570228 2. Department of Image Processing, Luster LightTec, Beijing 100097

记样本. 尤其是多标记情况下, 为了避免类别数目增加导致的已标记样本相对稀疏的情况, 需要大量已标记样本. 但是, 与之相对立的是已标记样本尤其是多标记问题中已标记样本一般都价格昂贵, 需要花费大量时间和人工才能获得. 为了解决该问题, 科研工作者提出了若干可利用未标记样本的半监督多标记学习方法. ML-LGC (Multi-label local and global consistency)[8]、SMSE (Semi-supervised algorithm for multi-label learning by solving a Sylvester equation)[9] 和 MASS (Multi-label semi-supervised learning)[10] 分别在图半监督学习和 Hinge 损失基础上添加能反映类间关系的正则项, 一定程度上克服了图半监督学习和半监督支持向量机应用于多标记学习时无视类别之间关系的问题. TML (Transductive multi-label learning)[11] 在隐马尔可夫模型下利用未标记样本, 不仅考虑了类间相关关系, 还考虑了类间互斥关系. Semi-supervised subspace learning[12] 通过组合无监督子空间表达方法和有监督多标记学习方法得到半监督多标记学习方法. iMLCU (Inductive multi-label classification with unlabeled data)[13] 在 Rank-SVM 基础上增加包含未标记样本的损失项和约束项, 提出了一种归纳 (Inductive) 多标记分类算法. CNMF (Multi-label learning method based on constrained non-negative matrix factorization)[14] 通过最大化样本特征距离矩阵与样本标签距离矩阵之间的相似度, 得到未标记样本的所有类属. Tram (Transductive multi-label classification)[15] 假设样本标签在样本特征流形面上足够光滑, 在此基础上得到类似于随机游走模型的多标记学习方法, 并在多个真实数据上验证了其学习效果.

与这些半监督多标记学习方法不同, 本文以样本特征集与样本标签集的规范化依赖性为基础, 提出了一种新的半监督多标记学习方法, 称为最大规范化依赖性多标记半监督学习方法 (Normalized dependence maximization multi-label semi-supervised learning method, DMMS). 变量依赖性的判断在统计学中已经具有完备的理论. 近年来, 因为独立分量分析 (Independent component analysis) 的需要, 提出了若干基于再生核希尔伯特空间的依赖性度量方法, 包括 Kernel constrained covariance (KCC)[16]、Hilbert-Schmidt independence criterion (HSIC)[17] 以及 Kernel generalised variance (KGV)[18] 等. 其中, HSIC 对于独立性的估计具有形式简单和收敛速度快等特点, 已经在近年被用于聚类分析[19] 和结构发现[20] 以及多标记维数约简[21] 且均取得了非常好的效果.

本文在规范化 HSIC 的基础上, 将已有标签作为约束, 利用所有样本, 包括已标记和未标记样本对特征集和标签集的规范化依赖性进行估计, 并以最大化该估计值作为优化目标, 最终通过求解带边界的迹比值问题为未标记数据打上标签. 理论上, DMMS 以基于统计理论的依赖性作为理论基础, 可以通过增加样本数目, 包括未标记样本数目提高依赖性估计的准确性; 另一方面, 无论样本同时属于多少个类, DMMS 都将该样本的标签组合看做标签集中一个点并映射至再生核希尔伯特空间, 因此 DMMS 是适用于多标记学习问题的直推 (Transductive) 半监督学习方法. 本文在多个真实多标记数据库与其他经典多标记学习方法, 包括 MLKNN 和 Binary relevance 以及其他半监督多标记学习方法, 包括 ML-LGC[8] 和 Tram[15] 做了对比实验. 实验表明 DMMS 可以有效从已标记和未标记样本中学习, 尤其是已标记样本相对稀少时, 学习效果在多项指标上都有显著提升.

## 1 HSIC 简介

HSIC 是一种基于核的独立性度量方法, 通过计算再生核希尔伯特空间上 Hilbert-Schmidt 互协方差算子范数得到独立性判断准则.

假设 $\mathcal{X}$ 和 $\mathcal{Y}$ 都是可分度量空间, 且 $\mathcal{F}$ 和 $\mathcal{G}$ 分别是 $\mathcal{X}$ 和 $\mathcal{Y}$ 的再生核希尔伯特空间. 记 $\mathcal{X}$ 到 $\mathcal{F}$ 上的映射为 $\Phi : \mathcal{X} \to \mathcal{F}$, 得到 $\mathcal{X}$ 上核函数为

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathcal{F}}, \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} \tag{1}$$

这里 $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ 表示空间 $\mathcal{F}$ 上内积. 类似地, 可以记 $\mathcal{Y}$ 到 $\mathcal{G}$ 的映射为 $\Psi : \mathcal{Y} \to \mathcal{G}$, 得到相应核函数为

$$l(\boldsymbol{y}, \boldsymbol{y}') = \langle \Psi(\boldsymbol{y}), \Psi(\boldsymbol{y}') \rangle_{\mathcal{G}}, \quad \boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y} \tag{2}$$

假设 $\Pr_{\mathcal{X}\times\mathcal{Y}}$ 是 $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$ 上的联合分布, $\Gamma$ 和 $\Lambda$ 分别是 $\mathcal{X}$ 和 $\mathcal{Y}$ 的 Borel 集. 相应的边缘分布分别记为 $\Pr_{\mathcal{X}}$ 和 $\Pr_{\mathcal{Y}}$, 互协方差算子 $C_{\boldsymbol{xy}} : \mathcal{G} \to \mathcal{F}$ 定义为

$$C_{\boldsymbol{xy}} = \mathrm{E}_{\boldsymbol{x},\boldsymbol{y}} \left[ \Phi(\boldsymbol{x}) \otimes \Psi(\boldsymbol{y}) \right] - \mu_{\boldsymbol{x}} \otimes \mu_{\boldsymbol{y}} \tag{3}$$

这里, $\mu_{\boldsymbol{x}}$ 和 $\mu_{\boldsymbol{y}}$ 分别表示 $\Phi(\boldsymbol{x})$ 和 $\Psi(\boldsymbol{y})$ 的期望. $\otimes$ 表示张量积, 对任意 $\boldsymbol{f} \in \mathcal{F}$ 和 $\boldsymbol{g} \in \mathcal{G}$, 有 $\boldsymbol{f} \otimes \boldsymbol{g} : \mathcal{G} \to \mathcal{F}$ 为

$$(\boldsymbol{f} \otimes \boldsymbol{g})\boldsymbol{h} = \boldsymbol{f} \langle \boldsymbol{g}, \boldsymbol{h} \rangle_{\mathcal{G}}, \qquad \forall \boldsymbol{h} \in \mathcal{G} \tag{4}$$

$C_{\boldsymbol{xy}}$ 可以看成 Hilbert-Schmidt 算子, 而所谓的 HSIC 即定义为 $C_{\boldsymbol{xy}}$ 的 Hilbert-Schmidt 算子范数, 也即:

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \Pr_{\mathcal{X}\times\mathcal{Y}}) = \|C_{\boldsymbol{xy}}\|^2_{\mathrm{HS}} \tag{5}$$

在观察得到数据 $Z = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ 的基础上, 可以给出 HSIC 的经验估计值:

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, Z) = (n-1)^{-2}\mathrm{tr}[HKHL] \tag{6}$$

其中, $\mathrm{tr}[\cdot]$ 表示求迹, $H = I - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$, $I$ 为单位矩阵, $\boldsymbol{e}$ 是元素值全为 1 的列向量, $K, L$ 分别是核 $k$ 和 $l$ 关于观测值 $Z$ 的 Gram 矩阵, 即 $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 以及 $L_{ij} = l(\boldsymbol{y}_i, \boldsymbol{y}_j)$. HSIC 的经验估计值在理论上已被证明具有收敛速度快以及计算简单等优点, 它的值越大说明 $\mathcal{X}$ 和 $\mathcal{Y}$ 关联性越强, 等于 0 时说明 $\mathcal{X}$ 和 $\mathcal{Y}$ 相互独立.

## 2　最大规范化依赖性多标记半监督学习方法

考虑到样本特征与其标签具有联系这一基本假设, 本文在 HSIC 的基础上对样本特征集与标签集之间的关联程度进行量化.

给定已标记数据集和未标记数据集分别为 $V = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y} | i = 1, \cdots, v\}$ 和 $U = \{(\boldsymbol{x}_j, \boldsymbol{y}_j) \in \mathcal{X} \times \mathcal{Y} | j = v+1, \cdots, v+u\}$, 其中 $\mathcal{X}$ 和 $\mathcal{Y}$ 分别是样本特征和标签所在空间. 记样本可能的类别总数为 $m$, 那么 $\mathcal{Y}$ 中标签向量是 $m$ 维列向量. 其中, 已标记样本 $\boldsymbol{x}_i$ $(i = 1, \cdots, v)$ 的类别已知, 相应标签向量 $\boldsymbol{y}_i$ 记为

$$y_{ki} = \begin{cases} 1, & \boldsymbol{x}_i \text{属于第} k \text{类} (1 \le k \le m) \\ -1, & \text{否则} \end{cases} \tag{7}$$

未标记样本 $\boldsymbol{x}_j$ $(j = v+1, \cdots, v+u)$ 的类别未知, 可以设定它的标签向量 $\boldsymbol{y}_j$ 为实向量, 当中元素表示置信值. 比如, $y_{kj}$ 是第 $j$ 个样本属于 $k$ 类的置信程度. DMMS 的目标即是求得这些置信值. 记

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{v+u}], \quad Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{v+u}] \tag{8}$$

给定 $\mathcal{X}$ 和 $\mathcal{Y}$ 上的核函数分别为 $k(\boldsymbol{x}, \boldsymbol{x}')$ 和 $l(\boldsymbol{y}, \boldsymbol{y}')$, 可以得到它们关于 $X$ 和 $Y$ 的 Gram 矩阵 $K$ 和 $L$, 从而有 $\|C_{\boldsymbol{xy}}\|^2_{\mathrm{HS}}$ 的估计值为

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, X, Y) = (n-1)^{-2}\mathrm{tr}[HKHL] \tag{9}$$

$\mathcal{F}$ 和 $\mathcal{G}$ 分别是 $\mathcal{X}$ 和 $\mathcal{Y}$ 的再生核希尔伯特空间, $H$ 的定义同式 (6), $n = v + u$ 表示样本总数. 简单起见, 标签集上的核函数取为线性核, 即 $l(\boldsymbol{y}, \boldsymbol{y}') = \boldsymbol{y}'^{\mathrm{T}}\boldsymbol{y} (\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y})$, 于是

$$\mathrm{tr}[HKHL] = \mathrm{tr}[HKHY^{\mathrm{T}}Y] = \mathrm{tr}[YHKHY^{\mathrm{T}}] \tag{10}$$

单纯采用式 (10) 的值作为样本特征集与标签集之间的关联性度量, 会受到标签集的尺度影响, $Y$ 的尺度越大它的值越大. 针对该问题, 对式 (10) 进行类似于 Blaschko 等的规范化[20], 得到以下优化目标:

$$\max_{Y} \frac{\mathrm{tr}\,[YHKHY^{\mathrm{T}}]}{\sqrt{\mathrm{tr}[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y]}} \tag{11}$$

在式 (11) 中增加常数项 $\sqrt{\mathrm{tr}[HKHK]}$ 可得到它的同解问题如下:

$$\max_{Y} \rho(X, Y) = \frac{\mathrm{tr}\,[YHKHY^{\mathrm{T}}]}{\sqrt{\mathrm{tr}[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y]\mathrm{tr}[HKHK]}} \tag{12}$$

上式分母因子项 $\mathrm{tr}[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y] = \mathrm{tr}[HLHL]$ 和 $\mathrm{tr}[HKHK]$ 分别是空间 $\mathcal{F}$ 和 $\mathcal{G}$ 到自身的 Hilbert-Schmidt 算子范数 $\|C_{\boldsymbol{yy}}\|^2_{\mathrm{HS}}$ 和 $\|C_{\boldsymbol{xx}}\|^2_{\mathrm{HS}}$ 的估计值. 因此, 某种意义上, 本文优化目标可以看成是样本特征集与标签集的相关系数的估计值. 可以证明, 若 $\rho(X, Y)$ 存在, 则有:

$$0 \le \rho(X, Y) \le 1 \tag{13}$$

先证 $\rho(X, Y) \le 1$. 事实上, 令 $\tilde{K} = HKH$, $\tilde{L} = HLH$, 由 $H^2 = H$ 可知

$$\mathrm{tr}[HKHL] = \mathrm{tr}[HKHHLH] = \left\langle \tilde{K}, \tilde{L} \right\rangle_F \tag{14}$$

这里 $\langle \cdot \rangle_F$ 表示 Frobenius 内积. 同理,

$$\mathrm{tr}[HKHK] = \left\langle \tilde{K}, \tilde{K} \right\rangle_F \tag{15}$$

$$\mathrm{tr}[HLHL] = \left\langle \tilde{L}, \tilde{L} \right\rangle_F \tag{16}$$

由施瓦茨不等式立刻就有:

$$\left\langle \tilde{K}, \tilde{L} \right\rangle^2_F \le \left\langle \tilde{K}, \tilde{K} \right\rangle_F \left\langle \tilde{L}, \tilde{L} \right\rangle_F \tag{17}$$

即 $\rho(X, Y) \le 1$. 下证 $0 \le \rho(X, Y)$. 记 $\tilde{Y} = YH$, $\tilde{Y}(i, \cdot)$ 表示 $\tilde{Y}$ 第 $i$ 行, 由 $K$ 是半正定矩阵可知,

$$\begin{aligned} \mathrm{tr}\,[HKHL] = \mathrm{tr}\,[YHKHY^{\mathrm{T}}] = \\ \sum_i \mathrm{tr}\left[\tilde{Y}(i, \cdot) K Y(i, \cdot)^{\mathrm{T}}\right] \ge 0 \end{aligned} \tag{18}$$

另一方面, 由内积定义可知 $\mathrm{tr}[HLHL] = \left\langle \tilde{L}, \tilde{L} \right\rangle_F \ge 0$, 且类似也有 $\mathrm{tr}[HKHK] \ge 0$. 因此, 如果 $\rho(X, Y)$ 存在, 则必有 $0 \le \rho(X, Y)$.

选用 $\rho(X, Y)$ 而非单纯的 $\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, X, Y)$ 作为优化目标, 可以避免 $Y$ 的尺度造成的无解问题. 最后, 注意到 $Y$ 对应于已标记样本部分事实上是已知的, 本文将这些已知标签作为边界条件. 记 $Y_V$ 和 $Y_U$ 分别为 $Y$ 对应于已标记和未标记样本部分,

DMMS 最终写为以下优化问题:

$$\begin{aligned}&\max_{Y}\ \frac{\mathrm{tr}\left[YHKHY^{\mathrm{T}}\right]}{\sqrt{\mathrm{tr}\left[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y\right]}}\\&\text{s.t.}\quad Y_V\equiv[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_v],\\&\qquad \|Y_U\|_F^2=\tau,\quad \tau\text{为常数}\end{aligned}\tag{19}$$

其中, $\|\cdot\|_F$ 是 Frobenius 范数, $\tau>0$ 是预先给定常数, 用于协调已标记和未标记数据, 避免 $Y_U$ 尺度过大减弱 $Y_V$ 对于比值大小的贡献率.

理论上, HSIC 估计值在样本数趋于无穷大时, 将以极大概率收敛至真实值[22]. DMMS 以 HSIC 估计值为算法基础, 选择与样本特征空间规范化依赖程度最高的标签预测值作为未标记样本的最终标签. 大量样本, 包括未标记样本的加入有助于提高此估计的准确度, 进而提高分类精度. 此外, DMMS 对于标签集中元素的维数没有限制, 因此 DMMS 是一种适用于多标记学习问题的直推半监督学习方法.

## 3 DMMS 的求解

首先给出式 (19) 具有最优解的证明. $Y$ 中的 $Y_V$ 部分非 0, 由 $\tilde{L}=HY^{\mathrm{T}}YH$ 且 $\mathrm{tr}[HLHL]=\left\langle\tilde{L},\tilde{L}\right\rangle_F$ 可知分母 $\mathrm{tr}[HLHL]\neq 0$. 又式 (19) 中的分子分母俱是关于 $Y_U$ 的连续函数, 因此式 (19) 的目标函数是关于 $Y_U$ 的连续函数, 在约束 $\|Y_U\|_F^2=\tau$ 下必有最大值.

接下来可以通过求解带边界的迹比值问题得到式 (19) 的最优解.

### 3.1 简化为迹比值问题

尽管式 (19) 在理论上存在最优解, 但是直接求解该问题很困难, 下面做些简化. Blaschko 等对类似于式 (19) 的最优化问题采用贪心策略[20], 每次只求取 $Y$ 中的一行 (即某一类), 从而近似得到整个问题的解. 将 $Y$ 视为一行, 有:

$$\begin{aligned}&\sqrt{\mathrm{tr}\left[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y\right]}=\\&\qquad\sqrt{\mathrm{tr}\left[YHY^{\mathrm{T}}YHY^{\mathrm{T}}\right]}=\mathrm{tr}\left[YHY^{\mathrm{T}}\right]\end{aligned}\tag{20}$$

本文考虑在 $Y$ 是矩阵的情况下直接用 $\mathrm{tr}\left[YHY^{\mathrm{T}}\right]$ 近似代替 $\sqrt{\mathrm{tr}\left[HY^{\mathrm{T}}YHY^{\mathrm{T}}Y\right]}$ 以达到简化以及加快计算的目的. 记 $A=HKH$, 简化后的优化问题为

$$\begin{aligned}&\max_{Y}\ \frac{\mathrm{tr}\left[YAY^{\mathrm{T}}\right]}{\mathrm{tr}\left[YHY^{\mathrm{T}}\right]}\\&\text{s.t.}\quad Y_V\equiv[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_v],\\&\qquad \|Y_U\|_F^2=\tau,\quad \tau\text{为常数}\end{aligned}\tag{21}$$

可以证明, 简化后优化问题依然存在最优解. 事实上, 由 $H$ 定义可知, $H$ 是半正定矩阵, 而

$$\mathrm{tr}\left[YHY^{\mathrm{T}}\right]=\sum_i Y(i,:)HY(i,:)^{\mathrm{T}}\tag{22}$$

又 $Y\neq 0$, 因此一般都有分母 $\mathrm{tr}\left[YHY^{\mathrm{T}}\right]\neq 0$, 这里 $Y(i,:)$ 是 $Y$ 第 $i$ 行. 接下来类似于式 (19) 有解证明, 立刻就有式 (21) 同样有最优解.

### 3.2 求解迹比值问题

广而言之, 式 (21) 可以看作迹比值 (Trace ratio) 问题[23−24]. 与其他迹比值问题相比, 因为含有边界条件和约束条件, 式 (21) 的求解会更为困难. 为了求解该问题, 首先将 $A$ 按照已标记和未标记样本的划分分成 4 部分:

$$A=\begin{bmatrix}A_V & A_{VU}\\ A_{UV} & A_U\end{bmatrix}\tag{23}$$

其中, $A_V$ 和 $A_U$ 分别对应已标记和未标记样本, $A_{UV}^{\mathrm{T}}=A_{VU}$. 类似的, 有:

$$H=\begin{bmatrix}H_V & H_{VU}\\ H_{UV} & H_U\end{bmatrix}\tag{24}$$

于是

$$\begin{aligned}\mathrm{tr}[YAY^{\mathrm{T}}]&=\mathrm{tr}[Y_VA_VY_V^{\mathrm{T}}+2Y_VA_{VU}Y_U^{\mathrm{T}}+Y_UA_UY_U^{\mathrm{T}}]\\ \mathrm{tr}[YHY^{\mathrm{T}}]&=\mathrm{tr}[Y_VH_VY_V^{\mathrm{T}}+2Y_VH_{VU}Y_U^{\mathrm{T}}+Y_UH_UY_U^{\mathrm{T}}]\end{aligned}\tag{25}$$

记

$$\begin{aligned}f(Y_U)&=\mathrm{tr}[YAY^{\mathrm{T}}]\\ g(Y_U)&=\mathrm{tr}[YHY^{\mathrm{T}}]\end{aligned}\tag{26}$$

与其他迹比值最优化问题求解方式[23−24] 类似, 本文希望在迭代过程中逐步提高迹的比值, 并且在若干迭代步后能够收敛至最优解. 对于给定 $Y_U^b(\|Y_U^b\|_F^2=\tau)$, 令

$$\lambda^b=\frac{f(Y_U^b)}{g(Y_U^b)}\tag{27}$$

有

$$f(Y_U^b)-\lambda^b g(Y_U^b)=0\tag{28}$$

令 $F(Y_U)=f(Y_U)-\lambda^b g(Y_U)$, 求得:

$$Y_U^*=\arg\max_{\|Y_U\|_F^2=\tau}F(Y_U)\tag{29}$$

于是

$$\begin{aligned} &F(Y_U^*) \geq F(Y_U^b) \\ &f(Y_U^*) - \lambda^b g(Y_U^*) \geq 0 \\ &\frac{f(Y_U^*)}{g(Y_U^*)} \geq \frac{f(Y_U^b)}{g(Y_U^b)} \end{aligned} \tag{30}$$

从式 (30) 可知, 上述步骤可以保证每次迭代过程中都找到新的 $Y_U^*$, 使得迹的比值只升不降. 假设式 (21) 的最优值点为 $Y_U^M$, 相应最优值为 $\lambda_M$, 迭代过程中迹的比值系列为

$$\lambda^0, \lambda^1, \cdots, \lambda^n, \lambda^{n+1}, \cdots \tag{31}$$

可以证明这个系列必将收敛到 $\lambda_M$. 事实上, 由

$$\begin{aligned} &\lambda^0 \leq \lambda^1 \leq \cdots \leq \lambda^n \leq \lambda^{n+1} \leq \cdots \\ &\lambda^i \leq \lambda_M (\forall i) \end{aligned} \tag{32}$$

可知这是一个单调上升且有上界的数列, 从而必有极限. 令 $\lambda_{\lim} = \lim_{i\to\infty} \lambda^i$, 下面反证必有 $\lambda_{\lim} \equiv \lambda_M$. 假设 $\lambda_{\lim} \neq \lambda_M$, 那么 $\lambda_{\lim} < \lambda_M$, 令

$$\lambda_d = \lambda_M - \lambda_{\lim} \tag{33}$$

那么对于任意 $i$ 均有 $\lambda_M - \lambda^i \geq \lambda_d$, 也即

$$f(Y_U^M) - \lambda^i g(Y_U^M) \geq \lambda_d g(Y_U^M) \tag{34}$$

另一方面, 由 $\lambda_{\lim} = \lim_{i\to\infty} \lambda^i$, 根据极限定义, 任意 $\varepsilon > 0$, 存在 $i_0$, 对任意 $i > i_0$ 有:

$$\lambda^{i+1} - \lambda^i < \varepsilon \tag{35}$$

也即

$$f(Y_U^{i+1}) - \lambda^i g(Y_U^{i+1}) < \varepsilon g(Y_U^{i+1}) \tag{36}$$

$g(Y_U)$ 是连续函数, 在有界闭集内 $g(Y_U^{i+1})$ 小于某常数, 又 $\lambda_d g(Y_U^M)$ 是一个常数, 综合式 (34) 和式 (36), 只要 $i_0$ 足够大, 就有:

$$f(Y_U^M) - \lambda^i g(Y_U^M) > f(Y_U^{i+1}) - \lambda^i g(Y_U^{i+1}) \tag{37}$$

这与 $Y_U^{i+1} = \arg\max_{\|Y_U\|_F^2=\tau} f(Y_U) - \lambda^i g(Y_U)$ 矛盾, 因此必有 $\lambda_{\lim} \equiv \lambda_M$, 相应达到该比值的点 $Y_U^{\lim}$ 即为最优值点.

求解式 (21) 的详细算法步骤总结在算法 1 中.

**算法 1. 求解迹比值最优化问题 (21)**

**输入.** 矩阵 $A$ 和 $H$, 已标记样本的标签矩阵 $Y_V$ 以及参数 $\tau$.

**输出.** 未标记样本的标签矩阵 $Y_U$.

**步骤 1.** 按照式 (26) 得到函数 $f$ 和 $g$, 给定阈值 $\kappa > 0$ (本文取为 $10^{-7}$) 为很小的数.

**步骤 2.** 给定 $\lambda^a$=0, 随机初始化 $Y_U^b$ 并规范化, 使得 $\mathrm{tr}\left[Y_U^b (Y_U^b)^{\mathrm{T}}\right] = \tau$, 记 $\lambda^b = f(Y_U^b)/g(Y_U^b)$.

**步骤 3.** 循环执行步骤 3.1 和步骤 3.2 直到 $\lambda^b - \lambda^a < \kappa$.

**步骤 3.1.** 令 $F(Y_U) = f(Y_U) - \lambda^b g(Y_U)$, 求解得到新的 $Y_U^b = \arg\max_{\|Y_U\|_F^2=\tau} F(Y_U)$.

**步骤 3.2.** 令 $\lambda^a = \lambda^b$, $\lambda^b = f(Y_U^b)/g(Y_U^b)$.

**步骤 4.** 输出 $Y_U = Y_U^b$. $Y_U^b$ 中每一列 $Y_U(:, j)$ ($j = 1, \cdots, u$) 的第 $i$ 个数表示第 $j$ 个样本属于第 $i$ 类的置信度.

### 3.3 求解迹比值问题中的子问题

在提高迹的比值过程中需要求解另一子优化问题, 即式 (29) (算法 1 中步骤 3.1). 现将该优化问题重新叙述如下:

$$\begin{aligned} &\max_{Y_U} F(Y_U) \\ &\text{s.t.} \quad \|Y_U\|_F^2 = \tau, \quad \tau\text{为常数} \end{aligned} \tag{38}$$

其中

$$\begin{aligned} F(Y_U) = f(Y_U) - \lambda^b g(Y_U) = \\ \mathrm{tr}[Y_U (A_U - \lambda^b H_U) Y_U^{\mathrm{T}}] + \\ 2\mathrm{tr}[Y_V (A_{VU} - \lambda^b H_{VU}) Y_U^{\mathrm{T}}] + \\ \mathrm{tr}[Y_V (A_V - \lambda^b H_V) Y_V^{\mathrm{T}}] \end{aligned} \tag{39}$$

上式最后一项 $\mathrm{tr}[Y_V (A_V - \lambda^b H_V) Y_V^{\mathrm{T}}]$ 相对于 $Y_U$ 是常数, 不影响最后的解, 将它舍弃并记

$$\begin{aligned} &M = (A_U - \lambda^b H_U) \\ &N = Y_V (A_{VU} - \lambda^b H_{VU}) \end{aligned} \tag{40}$$

得到式 (38) 的同解问题为

$$\begin{aligned} &\max_{Y_U} F(Y_U) = \mathrm{tr}[Y_U M Y_U^{\mathrm{T}}] + 2\mathrm{tr}[N Y_U^{\mathrm{T}}] \\ &\text{s.t.} \quad \|Y_U\|_F^2 = \tau, \quad \tau\text{为常数} \end{aligned} \tag{41}$$

显然, 问题 (41) 存在最优解. 引入乘子形式有:

$$L(Y_U|\alpha) = F(Y_U) - \alpha(\|Y_U\|_F^2 - \tau) \tag{42}$$

由 KKT 条件可知 $\alpha \neq 0$, 且 $\alpha$ 与最优值点 $Y_U^*$ 应满足如下方程组:

$$\begin{cases} \dfrac{1}{2}\dfrac{\partial L(Y_U|\alpha)}{\partial Y_U} = (M - \alpha I) Y_U^{\mathrm{T}} + N^{\mathrm{T}} = 0 \\ \dfrac{\partial L(Y_U|\alpha)}{\partial \alpha} = \|Y_U\|_F^2 - \tau = 0 \end{cases} \tag{43}$$

注意到 $M$ 是实对称矩阵, 可正交分解为 $M = P\Lambda P^{\mathrm{T}}$, $\Lambda$ 是以 $M$ 的所有特征值为对角元素的对角矩阵, $P$ 是相应的特征向量矩阵. 于是, 可重写方程组 (43) 中第一个方程为

$$Y_U^{\mathrm{T}} = -P(\Lambda - \alpha I)^{-1} P^{\mathrm{T}} N^{\mathrm{T}} \tag{44}$$

结合方程组 (43) 的第二个方程就有:

$$\mathrm{tr}[P^{\mathrm{T}}N^{\mathrm{T}}NP(\Lambda-\alpha I)^{-2}]=\tau \tag{45}$$

也即

$$\sum_{i=1}^{u}\frac{\widehat{N}_i}{(\alpha-\lambda_i)^2}=\tau \tag{46}$$

其中

$$\widehat{N}=P^{\mathrm{T}}N^{\mathrm{T}}NP \tag{47}$$

$\widehat{N}_i$ 与 $\lambda_i$ 分别是 $\widehat{N}$ 和 $\Lambda$ 对角线上的第 $i$ 个元素. 求解式 (46) 可以得到拉格朗日乘子 $\alpha$ 的值. 特别地, 当未标记样本数 $u$ 比较大且 $\tau$ 较小时, 上式也可以简化为

$$\sum_{i=1}^{u}\frac{\widehat{N}_i}{\alpha^2}=\tau \tag{48}$$

即

$$\alpha=\sqrt{\frac{\mathrm{tr}[\widehat{N}]}{\tau}}\quad(\text{舍弃负值}) \tag{49}$$

最后由式 (44) 可以得到最优值点 $Y_U^*$. 求解式 (41) 的算法步骤总结为算法 2.

**算法 2. 求解优化问题 (41)**

**输入.** 矩阵 $A$ 和 $H$, 已标记样本的标签矩阵 $Y_V$, 参数 $\tau$, 当前迹比值 $\lambda^b$.

**输出.** 优化问题 (41) 的最优解 $Y_U^*$.

**步骤 1.** 按照式 (23) 和式 (24) 对 $A$ 和 $H$ 进行划分, 并由式 (40) 得到矩阵 $M$ 和 $N$.

**步骤 2.** 计算 $M$ 的特征向量和特征值, 得到特征向量矩阵 $P$ 和特征值对角矩阵 $\Lambda$ (实验仅选取了特征值最大的 6 个量), 并按照式 (47) 计算出 $\widehat{N}$.

**步骤 3.** 求解式 (46) 或直接由式 (49) 得到拉格朗日乘子 $\alpha$.

**步骤 4.** 按照式 (44) 对每一个拉格朗日乘子计算相应 $Y_U$, 选择使得 $f(Y_U)/g(Y_U)$ 最大的 $Y_U$ 作为最优值点 $Y_U^*$, 输出 $Y_U^*$.

## 4 实验

实验评测按照一定的采样比率从数据集中随机抽取少部分 (保证每一个类至少一个样本) 作为训练集, 剩下的作为测试集. 而且, 为了避免随机抽样对实验结果的影响, 同样的采样率下每个实验均进行了 5 次, 最后的结果以这几次实验结果的均值和标准差形式给出. 本文在不同采样率下将 DMMS 与经典多标记学习方法在多个真实多标记数据集上做了对比实验, 讨论了 DMMS 的执行效率以及 DMMS 中参数对识别精度的影响情况.

### 4.1 参数设置

实验涉及 5 种多标记学习方法, 下面就各种方法当中的参数给予说明和设定:

1) DMMS: DMMS (Matlab 实现) 是本文提出的方法, 当中的参数 $\tau$ 取为 1, 样本特征集上的核函数选为常用的高斯核, 即

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right),\quad \boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{X} \tag{50}$$

参数 $\sigma$ 取为样本特征集上任意两点欧氏距离的平均值. DMMS 方法也可以选用别的核函数, 但是因为本文目的在于引入 DMMS 方法, 而高斯核已足够说明 DMMS 方法的有效性, 因此关于其他核函数的选择不在本文讨论范围.

2) MLKNN: MLKNN (Matlab 实现) 改进 $k$ 近邻法使之适用于多标记情况, 需要构建近邻图. 近邻数设定为 15, 并将式 (50) 作为构图过程中任意两个样本特征之间相似度 (距离) 的计算公式. 参数 Smooth 设定为 1.

3) Binary relevance (Java 实现): Binary relevance 为每类分别建立分类器, 将多标记学习问题转换成若干二分类问题, 基础分类器采用朴素贝叶斯方法.

4) ML-LGC (Matlab 实现): ML-LGC 是半监督多标记学习方法, 需要同时构建样本特征集和类别标签集上的近邻图. 样本特征集上近邻图的构建与 MLKNN 一致. 考虑到类别总数通常不会很大且测试集上样本标签未知, 类别标签集上近邻图的构建只针对训练集且构建的是全连接图. 根据实验结果以及 Zha 等对参数设置的讨论[8], 协调样本标签和样本特征两幅图的权重因子分别取为 0.1 和 1.

5) Tram: Tram (Matlab 实现) 是半监督多标记学习方法, 分为概念学习和预测两步. 概念学习步骤需要对样本特征构建近邻图, 构建该近邻图的参数设定与 DMMS 方法一致. 概念预测步骤采用直推方式.

### 4.2 数据集

实验所用数据集是多标记学习领域常用或标准数据集, 包括: Scene、Emotions、Bibtex 和 Reuters. Scene[25] 是语义场景分类数据集; Emotion[26] 是音乐情感分类数据集; Bibtex[27] 来源于 ECML/PKDD 2008, 是文本标注数据集; Reuters(Rcv1)[28] 是著名的文本分类标准数据集, 本文使用了当中的第一子集. 这些数据集均可从开源项目 mulan[29] 的主页 (http://mlkd.csd.auth.gr/multilabel.html) 下载得

到. 它们的具体情况已经列在了表 1 里. 其中, Name、Domain、Instances、Attributes 和 Class 分别表示数据库名称、所属领域、样本总数、样本特征维数和类别总数. Cardinality 表示每个样本平均所属的类别数, 而 Density 是 Cardinality 与类别总数的商值.

### 4.3 评测指标

传统的单标记分类问题中的评测指标, 包括准确率 (Accuracy)、查准率 (Precision)、查全率 (Recall) 和 F-measure 等都不适用于多标记学习问题. 多标记学习问题中的评测要比单标记学习的复杂很多. Schapire 等定义了目前多标记学习中的 5 种常用评价指标[30], 它们的具体公式可参见原文, 这里简介如下:

1) 汉明损失 (Hamming loss): 指定阈值后, 可以通过样本类属置信值预测得到任意未标记样本的类属, 比如 $y_{ji}$ 大于阈值, 则认为第 $i$ 个样本属于第 $j$ 类. 汉明损失衡量了预测结果与样本实际类属之间的不一致程度, 即样本属于某类但未被识别出, 或不属于某类却被误判的可能性.

2) 1-错误率 (One-error): 描述了对任一样本类属置信值最高的类属不是其实际类别的平均可能性, 在单标记学习中, 就演化成普通的分类错误率.

3) 覆盖率 (Coverage): 将任意样本对应的类属置信值降序排序, 覆盖率衡量了从置信值最高的类别开始, 平均需要跨越多少个类属才能覆盖样本所属的全部类别.

4) 排序损失 (Ranking loss): 表明了预测结果里, 真实所属类别的置信值低于非所属类别置信值的可能性.

5) 平均精度 (Average precision): 平均精度反映了置信值大于真实类别置信值的类属全是样本所属真实类别的可能性.

5 项指标值, 除了平均精度是越大越好之外 (最大为 1), 其余都是越小说明学习方法越有效.

### 4.4 实验结果与分析

本文首先在 4 个真实数据集上对比了 DMMS、MLKNN、Binary Relevance、ML-LGC 和 Tram 5 种多标记学习方法的学习效果. 所有数据集在实验前均进行了尺度归一, 使得各特征的期望为 0, 方差为 1. 实验结果总结为误差棒图 (均值 + 标准差), 其中横轴表示采样率 (采样率指已标记样本占总样本比率), 纵轴是各项指标值.

图 1 是 Scene 数据集上的结果. 从结果可知, DMMS 和 Tram 两种半监督方法总体要略好于其他学习方法, 这可能与半监督学习能够利用未标记数据有关. 进一步可以看到, 在所有采样率下 DMMS 方法的 1-错误率、覆盖率、排序损失和平均精度都要好于其他 4 种方法, 且已标记数据越稀少越明显. 比如采样率 0.01 时, 相比最接近 DMMS 的方法, DMMS 在平均精度和 1-错误率上的提高程度超过了 9 %, 排序损失、覆盖率则提高了 40 %以上. 汉明损失方面, 采样率小于 0.04 时, DMMS 要好于其他方法, 而采用率偏高时 DMMS 比其余学习方法略差. 计算汉明损失需要给定相应的阈值. DMMS 按照 Cardinality (定义见第 4.2 节) 给出的比例值, 将单个样本置信值所在区间分成两部分, 中间的临界点指定为该样本阈值. 其余方法如何设置阈值可参看原文.

图 2 是 Emotions 数据集上的结果. 总体上, DMMS、Tram 和 ML-LGC 这三种半监督学习方法的效果要好于其他方法. 具体而言, 当采样率大于 0.04 时, DMMS 和 Tram 这两种半监督学习方法在各项指标上的曲线几乎重合, 都要略好于其他学习方法. 采样率小于 0.04 时, 已标记数据严重不足, Tram 和 ML-LGC 两种半监督学习方法相对于其他监督学习方法而言并没有优势甚至略差, 这可能与 ML-LGC 需要利用已标记数据估计类间关系, 而 Tram 需要利用已标记数据估计先验概率有关. 至于 DMMS, 采样率偏小时, 虽然它的识别效果也有所下降, 但是相比于 MLKNN 和 Tram 而言, 下降的程度要低很多, 总体上依然要比其他监督学习方法要略好. 比如采样率为 0.01 时, 训练集几乎只包含了每一类中的一个样本, DMMS 的覆盖率、汉明损失、平均精度和排序损失相比于最接近它的方法分别提高了 6 %、9 %、5.9 % 和 10 %.

表 1　多标记学习数据集

Table 1　The datasets for multi-label learning

| Name | Domain | Instances | Attributes | Class | Cardinality | Density |
|---|---|---|---|---|---|---|
| Emotions | Music | 593 | 72 | 6 | 1.869 | 0.311 |
| Scene | Multimedia | 2 407 | 294 | 6 | 1.074 | 0.179 |
| Bibtex | Text | 7 395 | 1 836 | 159 | 2.402 | 0.015 |
| Reuters | Text | 6 000 | 47 236 | 101 | 2.880 | 0.029 |

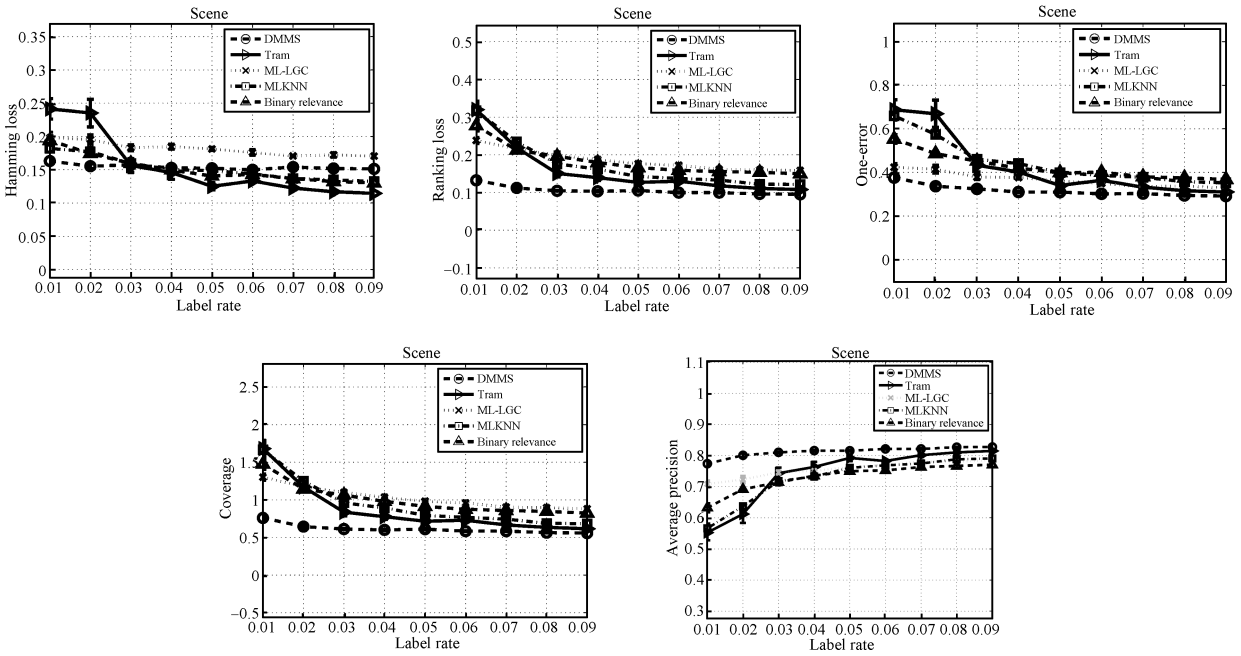



图 1 不同采样率下, Scene 数据集上的实验结果 (均值 ± 方差)

Fig. 1 Results on Scene data-set with different label rates (mean ± std)

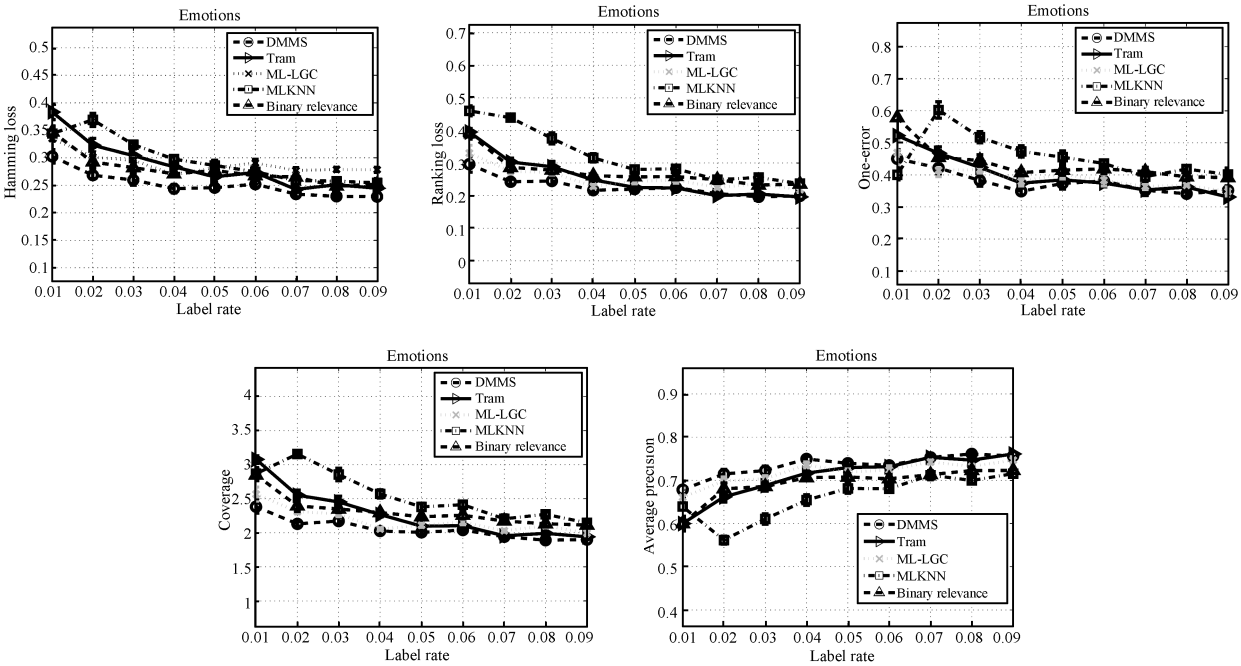



图 2 不同采样率下, Emotions 数据集上的实验结果 (均值 ± 方差)

Fig. 2 Results on Emotions data-set with different label rates (mean ± std)

图 3 是 Bibtex 数据集上的结果. Bibtex 数据集上样本数最少类别只含有 51 个样本, 因此设定采样率从 0.02 开始. 从图 3 可见, 汉明损失上 DMMS 与 MLKNN 曲线图几乎重合, 都要好于其他方法; 排序损失、覆盖率和平均精度上 DMMS 明显好于其他所有方法. 比如采样率为 0.02 时, 相较于最接近 DMMS 的方法, DMMS 在这三项指标上分别提升了 35 %、55 %和 30 %以上; 1-错误率上, DMMS 与 Binary relevance 接近, 都要好于其他方法.

图 4 是 Reuters 数据集上的结果. Reuters 数据集的样本维数比较高, 具有 47 236 维. 为了计算方便, 实验先用主成分分析方法将维数降

至 500 维. 同时, 为了保证采样率始于 0.01 且每类至少包含一个已标记样本, 实验只保留了样本数大于 100 的 48 个类别. 从实验结果可以看到, DMMS 的汉明损失与 MLKNN 几乎一致, 略好于其他方法. 其他指标上, 除了采样率 0.01 时, Tram 的 1-错误率值要好于其他方法之外, DMMS 在所有采样率上都要明显优于其他方法.

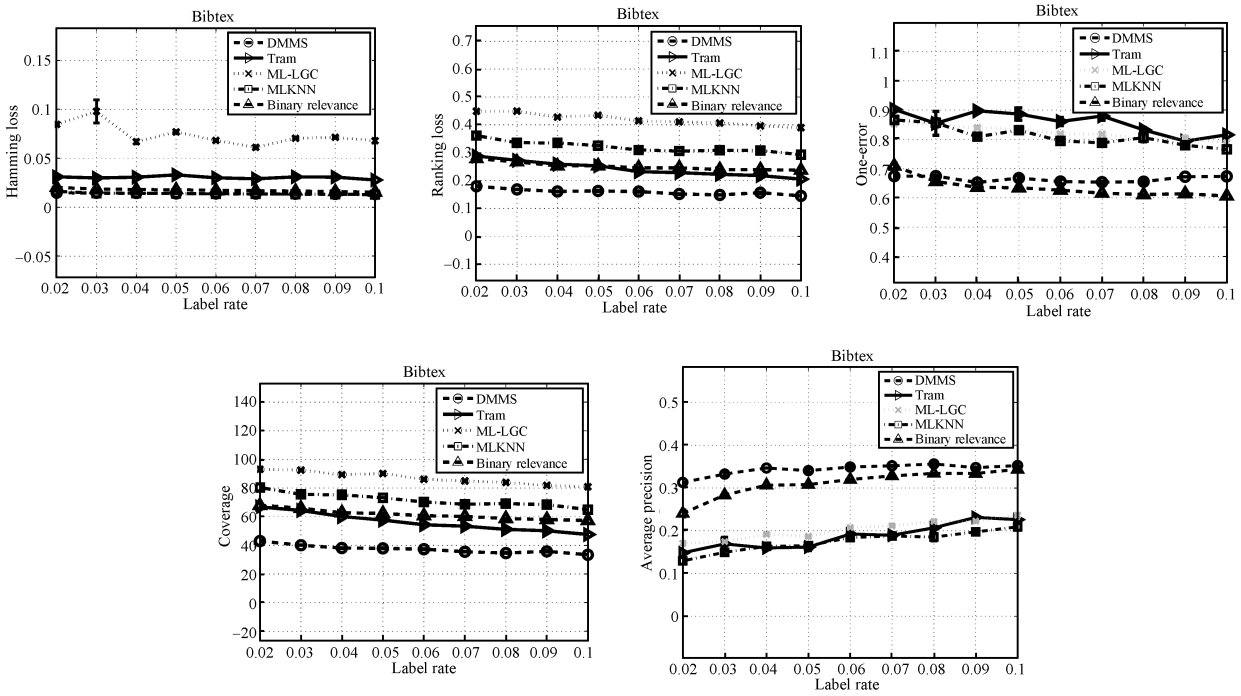



图 3　不同采样率下, Bibtex 数据集上的实验结果 (均值 ± 方差)

Fig. 3　Results on Bibtex data-set with different label rates (mean ± std)

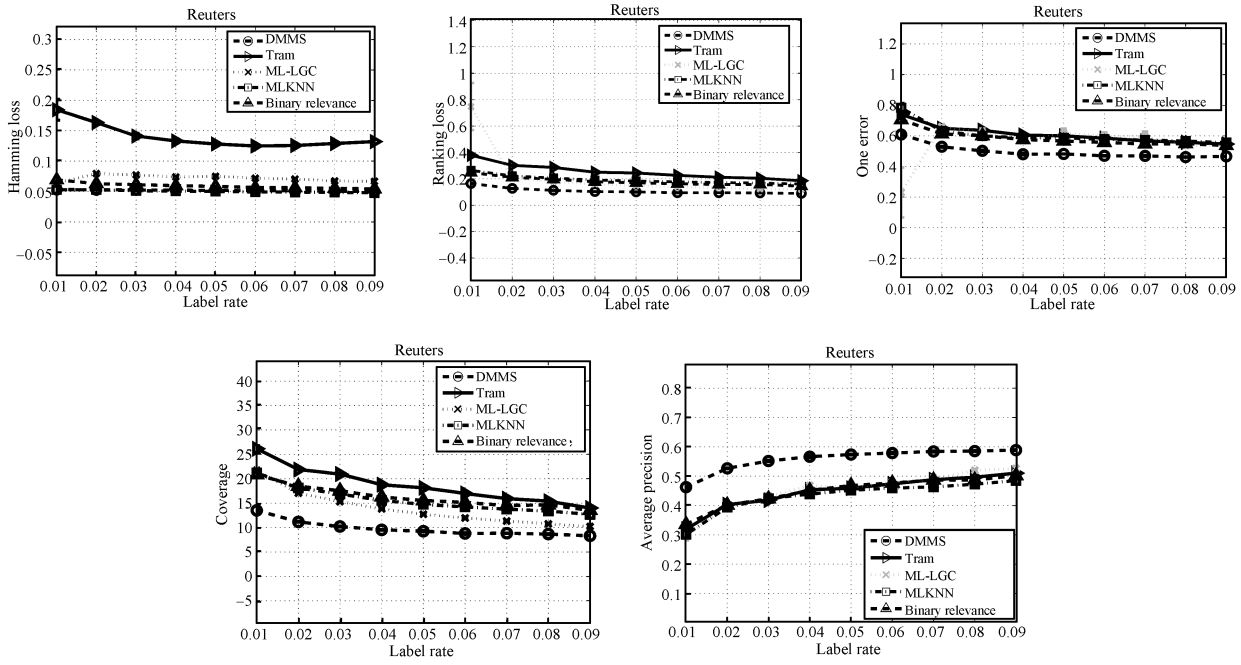



图 4　不同采样率下, Reuters 数据集上的实验结果 (均值 ± 方差)

Fig. 4　Results on Reuters data-set with different label rates (mean ± std)

进一步, 为了检验已标记样本稀少时 DMMS 是否显著优于其他方法, 本文选定 DMMS 作为控制方法 (Control method), 在显著性水平 $\alpha = 0.05$ 下将 DMMS 与其他方法作了 Holm 检验, 检验数据是采样率 0.04 以下 (包括 0.04) 各方法在所有数据集上的平均精度值, 检验结果参见表 2. 从表 2 可以看到, MLKNN、Tram、Binary relevance 和 ML-LGC 的 $p$ 值均小于修正后的显著水平 $\alpha/i$ 值, 因此应该拒绝零假设, 即认为 DMMS 显著优于其他方法.

各算法执行时间记录为所有采样率下算法单次执行所需平均时间, 总结上述实验执行情况为表 3. 从表 3 可知, Emotions 数据集上 DMMS 执行时间高于 ML-LGC、Tram 和 MLKNN. 尽管如此, DMMS 仅比时间花费最少的 MLKNN 方法慢了 0.244 秒. 其余数据上 DMMS 的执行效率仅次于 MLKNN, 要高于其他三种方法. 这充分说明了 DMMS 在执行效率上的有效性.

最后, 本文在限定采样率为 0.1 的情况下讨论了算法模型中参数 $\tau$ 对识别效果的影响, 总结为图 5. 图 5 中横轴取为 $\tau$ 的对数值, 对数底为 2, 纵轴设定为相对平均精度. 相对平均精度以 $\tau = 1$ 时, 平均精度为基准, 记为各点相对于 $\tau = 1$ 点时平均精度的变化率. 图 5 中绘制的线条代表了不同数据集的平均精度变化情况. 从图 5 可以看到, 尽管 DMMS 的平均精度会随着 $\tau$ 的增加而略有下降. 但是, 总体而言, $\tau$ 取值偏小时, 它对平均精度的影响甚微, 比如 $\tau$ 在图 5 中范围内, 各数据集上平均精度变化率都小于 0.04. 这说明在 $\tau$ 偏小时, DMMS 对 $\tau$ 的取值不敏感.

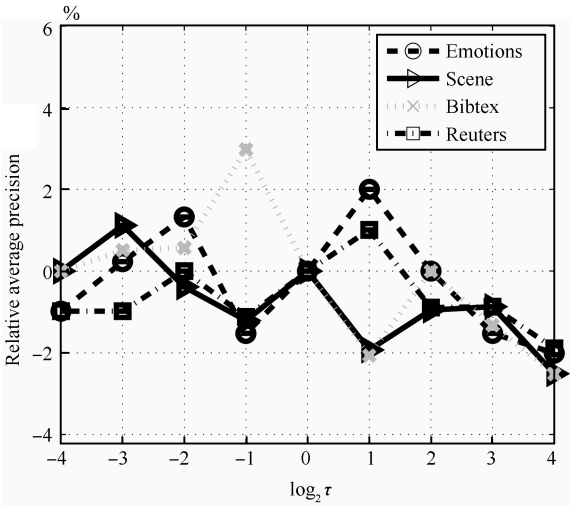



图 5 参数对相对平均精度 (%) 的影响

Fig. 5 The influence of parameter on relative average precision (%)

## 5 结论

DMMS 是在样本特征和样本标签规范化依赖性度量的基础上引入的一种新的直推半监督多标记学习方法. 多个真实数据集上的实验结果表明 DMMS 适用于多标记学习问题, 且在已标记样本稀少时, 可以利用未标记样本显著提高学习效果. 进一步, 我们将考虑在未来对 DMMS 进行扩展, 得到可归纳的半监督多标记学习方法.

表 2 Holm 检验结果 (DMMS 是控制方法)

Table 2 Holm test result (DMMS is the control method.)

| $i$ | 方法 | $z$ 值 | $p$ 值 | $\alpha/i$ | 是否拒绝零假设 |
|---|---|---|---|---|---|
| 4 | MLKNN | 5.925 | 0.000001 | 0.0125 | 拒绝 |
| 3 | Tram | 4.137 | 0.000035 | 0.0167 | 拒绝 |
| 2 | Binary relevance | 3.801 | 0.000144 | 0.0250 | 拒绝 |
| 1 | ML-LGC | 3.466 | 0.000528 | 0.0500 | 拒绝 |

表 3 各方法在不同数据集上的时间花费 (秒)

Table 3 The time cost of all methods on different datasets (s)

| | DMMS | ML-LGC | Tram | MLKNN | Binary relevance |
|---|---|---|---|---|---|
| Emotions | 0.467 | 0.400 | 0. 298 | 0.223 | 2.482 |
| Scene | 3.604 | 8.989 | 4.467 | 0.781 | 37.025 |
| Bibtex | 88.256 | 308.889 | 112.148 | 27.915 | 1 465.325 |
| Reuters | 38.808 | 154.807 | 43.538 | 8.206 | 297.581 |

## References


1 Tsoumakas G, Ioannis K, Ioannis V. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*. Berlin: Springer-Verlag, 2010. 667−685

2 Zhang M L, Zhou Z H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(8): 1819−1837

3 Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. *Machine Learning*, 2011, **85**(3): 333−359

4 Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 2011, **23**(7): 1079−1089

5 Tsoumakas G, Vlahavas I. Random k-labelsets: an ensemble method for multilabel classification. In: Proceedings of the 18th European Conference on Machine Learning. Warsaw, Poland: Springer, 2007. 406−417

6 Zhang M L, Zhou Z H. A k-nearest neighbor based algorithm for multi-label classification. In: Proceedings of the 2005 IEEE International Conference on Granular Computing. New York, USA: IEEE, 2005. 718−721

7 Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Proceedings of Advances in Neural Information Processing Systems. Cambridge, Massachusetts, USA: MIT Press, 2001. 681−687

8 Zha Z J, Mei T, Wang J D, Wang Z F, Hua X S. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 2009, **20**(2): 97−103

9 Chen G, Song Y Q, Wang F, Zhang C S. Semi-supervised multi-label learning by solving a Sylvester equation. In: Proceedings of the 2008 SIAM International Conference on Data Mining. Atlanta, USA: Curran Associates, 2008. 410−419

10 Li Yu-Feng, Huang Sheng-Jun, Zhou Zhi-Hua. Regularized semi-supervised multi-label learning. *Journal of Computer Research and Development*, 2012, **49**(6): 1272−1278 (李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习方法. 计算机研究与发展, 2012, **49**(6): 1272−1278)

11 Wang J D, Zhao Y H, Wu X Q, Hua X S. A transductive multi-label learning approach for video concept detection. *Pattern Recognition*, 2011, **44**(10−11): 2274−2286

12 Guo Y H, Schuurmans D. Semi-supervised multi-label classification. In: Proceedings of the 2012 European Conference, Machine Learning and Knowledge Discovery in Databases. Bristol, UK: Springer, 2012. 355−370

13 Wu L, Zhang M L. Multi-Label classification with unlabeled data: an inductive approach. In: Proceedings of the 2013 Asian Conference on Machine Learning. Cambridge, Massachusetts, USA: MIT Press/JMLR, 2013. 197−212

14 Liu Y, Jin R, Yang L. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the 21st National Conference on Artificial Intelligence. California, USA: AAAI Press, 2006. 421−426

15 Kong X N, Ng M K, Zhou Z H. Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering*, 2013, **25**(3): 704−719

16 Gretton A, Smola A, Bousquet O, Herbrich R, Belitski A, Augath M, Murayama Y, Pauls J, Schölkopf B, Logothetis N. Kernel constrained covariance for dependence measurement. In: Proceedings of 10th International Workshop on Artificial Intelligence and Statistics. New Jersey, USA: Society for Artificial Intelligence and Statistics, 2005. 12−23

17 Gretton A, Bousquet B, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of 16th International Conference on Algorithmic Learning Theory. Singapore: Springer, 2005. 63−77

18 Bach F R, Jordan M I. Kernel independent component analysis. *Journal of Machine Learning Research*, 2002, **3**: 1−48

19 Song L, Smola A, Gretton A, Borgwardt K M. A dependence maximization view of clustering. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 815−822

20 Blaschko M, Gretton A. Learning taxonomies by dependence maximization. In: Proceedings of Advances in Neural Information Processing Systems. Cambridge, Massachusetts, USA: MIT Press, 2008. 153−160

21 Zhang Y, Zhou Z H. Multi-label dimensionality reduction via dependency maximization. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence. California, USA: AAAI Press, 2008. 1503−1505

22 Gretton A, Fukumizu K, Teo C H, L. Song, Schölkopf B, Smola A J. A kernel statistical test of independence. In: Proceedings of Advances in Neural Information Processing Systems. Cambridge, Massachusetts, USA: MIT Press, 2008. 582−592

23 Jia Y Q, Nie F P, Zhang C S. Trace ratio problem revisited. *IEEE Transactions on Neural Networks*, 2009, **20**(4): 729−735

24 Wang H, Yan S C, Xu D, Tang X O, Huang T. Trace ratio vs. ratio trace for dimensionality reduction. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN: IEEE, 2007. 1−8

25 Boutell M R, Luo J B, Shen X P, Brown C M. Learning multi-label scene classification. *Pattern Recognition*, 2004, **37**(9): 1757−1771

26 Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I P. Multi-label classification of music into emotions. In: Proceedings of the 9th International Conference on Music Information Retrieval. Philadelphia, USA: Drexel University, 2008. 325−330

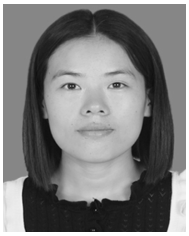
27 Katakis I, Tsoumakas G, Vlahavas I. Multilabel text classification for automated tag suggestion. In: Proceedings of the ECML/PKDD 2008 Discovery Challenge. Heidelberg, Berlin: Springer, 2008. 75−83

28 Lewis D D, Yang Y M, Rose T G, Li F. RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 2004, **5**: 361−397

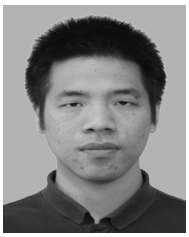
29 Tsoumakas G, Vilcek J, Xioufis E S. Mulan: a java library for multi-label learning [Online], available: http://mulan.sourceforge.net/datasets.html, January 1, 2010

30 Schapire R E, Singer Y. Boostexter: a boosting-based system for text categorization. *Machine Learning*, 2000, **39**(2−3): 135−168

**张晨光** 海南大学信息科学技术学院讲师. 2009 年获得北京工业大学硕士学位. 主要研究方向为图像处理, 模式识别.
E-mail: huzcg@foxmail.com
(**ZHANG Chen-Guang** Lecturer at the College of Information Science and Technology, Hainan University. He received his master degree from Beijing University of Technology in 2009. His research interest covers pattern recognition and image processing.)



**张 燕** 海南大学信息科学技术学院讲师. 主要研究方向为数据分析和数据挖掘. 本文通信作者.
E-mail: zhangyanouc@sina.com
(**ZHANG Yan** Lecturer at the College of Information Science and Technology, Hainan University. Her research interest covers data analysis and data mining. Corresponding author of this paper.)



**张夏欢** 北京凌云光视公司图像处理部图像算法工程师. 主要研究方向为图像处理和模式识别.
E-mail: zhanggongzi@yahoo.cn
(**ZHANG Xia-Huan** Image algorithm engineer in the Department of Image Processing, Luster LightTec. His research interest covers image processing and pattern recognition.)