一种基于正则化的半监督多标记学习算法

李宇峰 黄圣君 周志华
南京大学计算机软件新技术国家重点实验室，南京 210093
{liyf, huangsj, zhouzh}@lamda.nju.edu.cn

摘要 本文研究半监督多标记学习问题，其中每个训练样本可隶属于多个类标记并且大部分训练样本是未标记的。这类学习问题广泛存在于多标记学习的实际应用中，然而现有的大多数多标记学习算法只适用于传统监督学习框架。为解决该学习问题，本文提出了一种基于正则化的半监督多标记学习方法——MASS 方法，试图利用丰富的未标记样本与标记的关系信息以帮助提高性能。具体而言，在最小化经验风险的基础上，MASS 方法引入两种正则项分别用于描述多个标记的共性信息和约束相似样本应拥有相似的结构化多标记输出。本文提供一个快速交替优化算法取得整体凸优化问题的全局最优解。在网页分类和基因功能分析问题上的实验结果验证了本文方法的有效性。

关键词 **多标记学习；半监督学习**
中图法分类号 TP391

Regularized Semi-Supervised Multi-Label Learning

Li Yu-Feng, Huang Sheng-Jun, Zhou Zhi-Hua
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093

Abstract In this paper, we study the semi-supervised multi-label problem where each training example could be associated with multiple class labels while most of training examples are unlabeled. This scenario occurs in many real-world multi-label applications while most previous studies on multi-label learning typically work on traditional supervised learning setting. To address this problem, we propose a regularized Multi-lAbel Semi-Supervised learning (MASS) method which exploits both the abundant unlabeled examples and the label relationships simultaneously to help improve the performance. Specifically, besides the empirical risk minimization, MASS employs two regularizers to characterize the commonness among multiple labels and enforce similar instances share with similar structural multi-label outputs. Moreover, we use an efficient alternative method to obtain the optimal solution for the overall convex optimization problem. Experimental results on web page categorization and gene functional analysis demonstrate the effectiveness of our method.

Keywords multi-label learning; semi-supervised learning

1. 引言

多标记学习问题是机器学习领域中一类重要的学习问题。在该学习问题中，每个训练样本可隶属于多个不同的类标记，学习任务是为每个未见样本输出一个标记集合，其中集合大小是未知的。近年来，多标记学习框架在许多实际问题中取得了广泛的应用[17,19,9,4]。例如，网页分类通常是个多标记学习问题，其中每个网页往往对应于多个主题[17,19]；基因组功能分析是个多标记学习问题，其中每个基因可具有多个不同的功能，如新陈代谢、转录及蛋白质合成等[9]。

在多标记学习的实际应用中，大部分训练样本是没有标记的。例如，在网页分类问题中，网页是海量存在的，而网页真实标记的获取往往需要耗费大量人力资源；在基因组功能分析问题中，潜在的基因数目是海量的，而获取基因功能的真实标记需要重复多次科学实验甚至是难以实现的。此外，标记之间是有其内在关系的，否则多个不同的类标记不应该同时关联到同一个样本。因此，我们有必要研究当只有少量标记样本可用时，如何利用丰富的未标记样本及标记的

收稿日期：20xx-xx-xx
基金项目：本文得到国家自然科学基金（61073097, 60721002）、江苏省自然科学基金（BK2008018）、国家重点基础研究发展计划（2010CB327903）的资助

关系信息以提高泛化性能。现有的大多数多标记学习算法只适用于传统全监督学习框架，目前对该问题的研究只有少量的研究工作[15,7,20,25]。

本文提出了一种基于正则化的半监督多标记 MASS (Multi-lAbel Semi-Supervised) 方法。在传统经验风险最小化原理的基础上，MASS 方法引入两种正则项分别用于描述多个标记间的共性信息及约束相似的样本应拥有相似的结构化多标记输出。本文提供了一个快速交替优化算法取得整体凸优化问题的全局最优解。在网页分类和基因功能分析的实验结果表明，本文方法得到更好的性能。

本文余下部分组织如下：第 2 节简要介绍相关工作；第 3，4 节给出 MASS 方法及其快速优化算法；第 5 节给出实验结果；最后第 6 节总结全文。

2. 相关工作

现有的多标记学习方法集中于传统监督学习框架。具体做法可大致分为三类。第一类是将多标记学习问题转化为多个两类分类问题[14]，其中每个标记对应于一个两类分类问题，这种方法在类别标记较少、样本丰富时效果较好，但是在标记较多时，会遇到样本稀疏的问题，并且由于忽略了标记的关系信息，通常性能不佳。第二类是将多标记学习问题转化成标记排序问题[5,6,9]，这种方法侧重于标记之间序的正确性，但它需要额外学习一个阈值函数以得到最终的相关标记集合，而对这个阈值函数的学习本身就是一个困难的问题。第三类是将多标记学习问题与标记之间的结构信息进行结合[17,21,11,2]，这种方法在结构信息利用得当时，可以取得优异的性能，但在缺乏领域知识指导的情况下，几乎无法知道结构信息应该如何利用为好。

最近，有少量研究者开始研究直推式多标记学习(Transductive Multi-Label Learning)方法 [15,7,20,25]。Liu 等人[15]基于样本之间的相似性可由样本标记之间的相似性表出的假设，提出了 CNMF (Constrained Non-negative Matrix Factorization)方法对未标记样本的标记进行学习使得以上两种相似性差值最小。Chen 等人[7]基于相似度度量为样本和标记分别构建一个图，然后采用标记传播的思想对未标记样本的标记进行学习，整个优化问题可采用 Sylvester 方程进行快速求解。当标记样本很少时，基于少量标记估计得到的相似度度量可能是不可依赖的。Sun 等人[20]和孔等人[25]考虑多标记学习中的弱标记问题，即当标记数目很大时，训练样本的标记集合通常是不完备的。他们采用标记传播的思想对缺失标记进行学习。以上方法都是直推式方法，不能直接对除测试样本以外的未见样本进行预测。不同的是，本文 MASS 方法是一种归纳式多标记学习方法(Inductive Multi-Label Learning)，可以预测任意未见样本。

3. MASS 方法

给定数据集$\{\{x_i,Y_i\}_{i=1}^n,\{x_j\}_{j=n+1}^{n+m}\}$，其中$\{\{x_i,Y_i\}_{i=1}^n\}$表示标记样本，$\{\{x_j\}_{i=n+1}^{n+m}\}$表示未标记样本，$Y_i\in\{-1,+1\}^T$表示第$i$个训练样本的标记，$T$是标记的个数，则半监督多标记学习的任务是为该数据集学习一个映射函数$\boldsymbol{f}:\mathcal{X}\to 2^{\mathcal{Y}}$其中$\mathcal{X}$表示样本空间、$\mathcal{Y}$表示标记空间。

为了便于讨论，这里假定函数$\boldsymbol{f}$可分解为$\boldsymbol{f}=(f_1,f_2,\cdots,f_T)$其中每个$f_t$对应于一个标记。对于每个未见样本$x$，它的预测标记表示为$(\mathrm{sgn}(f_1(x),\ldots,\mathrm{sgn}(f_T(x))$其中sgn表示符号函数。不失一般性，假定每个$f_t$为线性模型，即$f_t(x)=\langle\boldsymbol{w}_t,\phi(x)\rangle$其中$\phi(x)$表示由核函数$k$导出的一个特征映射函数，即$k(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle$。这里$\langle\cdot,\cdot\rangle$表示由$k$张成的再生核空间(RHKS)$\mathcal{H}$中的内积。根据文献[10]，可证明$\boldsymbol{f}$同样属于一个再生核空间(RHKS)$\mathcal{H}_v$，其中的每个函数是一个向量函数。

本文考虑找到一个函数$\boldsymbol{f}$使得如下泛函最小化:

$$\begin{aligned} f^* &= \arg\min_{\boldsymbol{f}\in\mathcal{H}_v}\frac{1}{nT}\sum_{i=1}^{n}\mathbf{V}(x_i,Y_i,\boldsymbol{f}) \\ &+ \frac{\gamma_s}{T}\Omega(\boldsymbol{f})+\frac{\gamma_a}{(n+m)^2T^2}I(\boldsymbol{f}), \end{aligned} \tag{1}$$

其中$V(x_i,Y_i,\boldsymbol{f})$表示经验风险损失函数、$\Omega(\boldsymbol{f})$表示函数$\boldsymbol{f}$的正则项，通常用于控制函数$\boldsymbol{f}$的复杂度、$I(\boldsymbol{f})$表示流形正则项，要求相似的样本拥有相似的结构化多标记输出，γ_s和γ_a为正则化系数，用于折中这三项的重要性。在多标记学习算法中，经验风险损失函数经常采用 Hamming 损失函数[4,19,20,22,23]，即$V(x_i,Y_i,\boldsymbol{f})=\sum_{t=1}^{T}l(x_i,Y_{it},f_t)$其中$Y_{it}$表示$Y_i$的第$t$个元素，$l()$为两类问题的损失函数。本文采用支持向量机(SVM)中的 Hinge 损失函数作为$l()$的定义。关于$\Omega(\boldsymbol{f})$与$I(\boldsymbol{f})$的定义将分别由 3.1 节和 3.2 节给出。

3.1 正则项$\Omega(\boldsymbol{f})$描述标记间的共性

$\Omega(\boldsymbol{f})$的一种直接定义是为每个标记单独设定正则项，然后进行累加，例如$\Omega(\boldsymbol{f})=\sum_{t=1}^{T}\|\mathbf{w}_t\|^2$。这种定义忽略了标记关系，然而在许多实际任务中，标记之间应该是有内在相关性的。基于 Evgeniou 等人的工作[10]，本文假设所有的$\boldsymbol{w}_t$来自一个高斯分布其中$\boldsymbol{w}_0=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t$为该分布的均值，定义$\Omega(\boldsymbol{f})$如下：

$$\Omega(\boldsymbol{f}) = \mu\|\boldsymbol{w}_0\|^2 + \frac{\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^2}{T}, \quad (2)$$

正则化系数$\mu \in [-1, 0]$用于折中多个标记的共性项$\|\boldsymbol{w}_0\|^2$与特性项$\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^2$之间的重要性。其中，当$\mu$趋于-1 时，最小化$\Omega(\boldsymbol{f})$将使得所有$\boldsymbol{w}_t$趋于一致；而当$\mu$趋于 0 时，最小化$\Omega(\boldsymbol{f})$将使得所有$\boldsymbol{w}_t$趋于独立。

3.2 流形正则项$I(\boldsymbol{f})$要求相似样本拥有相似的结构化多标记输出

观察到多标记学习的实际应用中，相似的样本往往具有相似的结构化多标记输出。例如在网页分类中，主题之间通常具有结构性，相似的文档往往拥有相似的语义信息，它们的主题标记根据主题之间的结构性看来应是相似的。基于 Belkin 等人的工作[3]，本文采用流形正则项$I(\boldsymbol{f})$要求相似的样本拥有相似的结构化多标记输出。一个直接做法是通过简单累加多个两类流形正则项计算得到，

$$I(\boldsymbol{f}) = \sum_{i,j=1}^{n+m}(\boldsymbol{f}(x_i) - \boldsymbol{f}(x_j)'(\boldsymbol{f}(x_i) - \boldsymbol{f}(x_j)W_{ij} \quad (3)$$

其中$'$表示向量或矩阵的转置，$W \in \mathcal{R}^{(n+m)\times(n+m)}$表示样本间的相似度矩阵。注意到式(3)忽略了标记关系信息，本文想法是通过引入一个投影矩阵$P \in \mathcal{R}^{T\times T}$将标记投影一个新空间，在新空间里计算多标记输出的差异性。具体而言，引入矩阵$B^{-1} = P'P \in \mathcal{R}^{T\times T}$，则$I(\boldsymbol{f})$的数学表达为下式所示：

$$\begin{aligned} I(\boldsymbol{f}) &= \sum_{i,j=1}^{n+m}(P\boldsymbol{f}(x_i) - P\boldsymbol{f}(x_j))'(P\boldsymbol{f}(x_i) - P\boldsymbol{f}(x_j))W_{ij} \\ &= \sum_{i,j=1}^{n+m}(\boldsymbol{f}(x_i) - \boldsymbol{f}(x_j))'B^{-1}(\boldsymbol{f}(x_i) - \boldsymbol{f}(x_j))W_{ij} \\ &= \langle\hat{\boldsymbol{f}}'L\hat{\boldsymbol{f}}, B^{-1}\rangle \end{aligned} \quad (4)$$

其中$\hat{\boldsymbol{f}}$为$(n+m)\times T$大小的矩阵，每个元素为$\hat{\boldsymbol{f}}_{it} = f_t(x_i)$，$L = D - W$为$W$的 Laplacian 矩阵，其中$D$为对角矩阵，对角元素为$D_{ii} = \sum_{j=1}^{n+m}W_{ij}$，$i = 1, \ldots, n+m$。$\langle M, N\rangle$表示矩阵内积运算，即矩阵$M'N$的迹，常记为$trace(M'N)$。

这里矩阵B起到了度量多标记输出的作用。当领域知识充分时，B 矩阵可由领域知识给出，不需要学习。值得一提的是，当缺乏领域知识指导时，比如标记样本有限时，B矩阵的定义可能是不可依赖的。这种情况下一个可能的方式是结合大量的未标记样本对B矩阵进行学习，本文第 4 节给出了一种做法。

4 优化算法

由矩阵B的定义可知矩阵B为半正定阵，不失一般性，假定矩阵B的迹为1，即$B \in \mathcal{A}$其中$\mathcal{A} = \{B|B \succeq 0, trace(B) = 1\}$。将式(2)、式(4)代入式(1)中，可得如下优化问题，

$$\begin{aligned} (\boldsymbol{f}^*, B^*) &= \arg\min_{\boldsymbol{f}\in\mathcal{H}_v, B\in\mathcal{A}} \frac{1}{nT}\sum_{i=1}^{n}\mathbf{V}(x_i, Y_i, \boldsymbol{f}) \\ &+ \frac{\gamma_s}{T^2}\sum_{i=1}^{T}||\boldsymbol{w}_t||^2 + \frac{\mu\gamma_s}{T}||\boldsymbol{w}_0||^2 \\ &+ \frac{\gamma_a}{(n+m)^2T^2}\langle\hat{\boldsymbol{f}}^T L\hat{\boldsymbol{f}}, B^{-1}\rangle. \end{aligned} \quad (5)$$

根据表示定理[18]，可得如下定理，

定理 1 对任意半正定矩阵B，式(5)的最优解$\boldsymbol{f}$满足如下展式，

$$\boldsymbol{f}_t(x) = \sum_{i=1}^{n+m}\alpha_{it}k(x_i, x) \quad t = 1, \cdots, T \quad (6)$$

证明：结合文献[10]与[18]的工作可证明定理 1。这里限于篇幅，略去。

由定理 1，式(5)可重写为如下二次规划问题，

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\xi}} \quad & J_0^B(\boldsymbol{\alpha}, \boldsymbol{\xi}) = \frac{1}{nT}\boldsymbol{\xi}'\mathbf{1} + \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha} \\ \text{s.t.} \quad & Y_{it}(\boldsymbol{\alpha}_t^T K_i + b_t) \geq 1 - \xi_{it} \\ & i \in \{1, \cdots, m\}, t \in \{1, \cdots, T\} \\ & \boldsymbol{\xi} \geq \mathbf{0}, \end{aligned} \quad (7)$$

其中$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \cdots, \boldsymbol{\alpha}_T')'$，$\boldsymbol{\alpha}_t = (\alpha_{1t}, \cdots, \alpha_{(n+m)t})'$，$\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \cdots, \boldsymbol{\xi}_T')'$，$\boldsymbol{\xi}_t = (\xi_{1t}, \cdots, \xi_{(n+m)t})'$，$\mathbf{1}$为全1向量，$\boldsymbol{Q} = 2R \otimes K + \frac{2\gamma_a}{(n+m)^2T^2}B^{-1} \otimes KLK$其中$K$为 Gram 核矩阵，$R = \frac{\gamma_s}{T^2}I + \frac{\mu\gamma_s}{T^3}E$，$E$为全 1 矩阵，$\otimes$表示矩阵的 Kronecker 内积。注意到$K$、$R$、$L$、$B$都是半正定阵，根据半正定阵的 Kronecker 内积仍是半正定阵的原理，可知Q是半正定阵，进而可知式(7)是个凸优化问题。进一步根据文献[1]的工作可知，式(5)对于$(\boldsymbol{f}, B)$是个联合凸优化(Jointly Convex optimization)问题，可采用交替下降法求解。

4.1 固定B求解$\boldsymbol{\alpha}, \boldsymbol{\xi}$的快速算法

式(7)是个大规模的二次规划问题，难以对其直接求解。观察到式(7)中的约束矩阵为分块矩阵，即不同的$\{\boldsymbol{\alpha}_t, \boldsymbol{\xi}_t\}$不同时出现在约束中，因此本文采用块交替下降法(block-wise alternating descend method)逐次优化$\{\boldsymbol{\alpha}_t, \boldsymbol{\xi}_t\}$直至收敛[16]。根据文献[16]的工作，块交替下降法将以超线性速率收敛到最优解。这里每个子问题只需要涉及$\{\boldsymbol{\alpha}_t, \boldsymbol{\xi}_t\}$的求解，远远小于直接求解

表 1 网页分类数据 15%标记训练样本的结果，↓(↑)分别表示该指标值越小（大）越好

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	0.044±0.014	0.050±0.017	0.050±0.015	0.047±0.015	N/A	0.045±0.014	**0.042±0.014**
One-Error ↓	**0.460±0.139**	0.524±0.158	0.579±0.180	0.535±0.173	0.646±0.121	0.501±0.161	**0.453±0.141**
Coverage ↓	**4.392±1.357**	5.699±1.698	5.678±1.688	5.724±1.774	10.313±2.498	4.540±1.393	7.658±2.031
Ranking Loss ↓	**0.107±0.042**	0.145±0.050	N/A	0.150±0.063	0.280±0.085	0.115±0.045	0.194±0.056
Average Precision ↑	**0.631±0.106**	0.568±0.120	0.532±0.136	0.568±0.128	0.442±0.103	0.598±0.120	0.607±0.110

表 2 网页分类数据 10%标记训练样本的结果

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	0.045±0.015	0.051±0.017	0.054±0.017	0.047±0.015	N/A	0.046±0.015	**0.044±0.014**
One-Error ↓	**0.483±0.145**	0.548±0.165	0.607±0.183	0.562±0.183	0.652±0.111	0.518±0.166	**0.478±0.147**
Coverage ↓	**4.581±1.415**	5.988±1.766	6.202±1.822	5.822±1.664	11.455±2.715	4.767±1.467	7.752±2.137
Ranking Loss ↓	**0.113±0.044**	0.155±0.052	N/A	0.153±0.061	0.315±0.081	0.122±0.046	0.198±0.059
Average Precision ↑	**0.614±0.111**	0.547±0.126	0.507±0.140	0.550±0.133	0.423±0.093	0.584±0.125	0.590±0.115

表 3 网页分类数据 5%标记训练样本的结果

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	0.046±0.015	0.053±0.018	0.059±0.019	0.049±0.016	N/A	0.046±0.015	**0.045±0.014**
One-Error ↓	**0.518±0.160**	0.588±0.178	0.637±0.188	0.586±0.190	0.690±0.107	0.541±0.176	**0.512±0.162**
Coverage ↓	**4.941±1.451**	6.301±1.874	7.043±2.034	6.322±1.949	13.372±2.941	5.182±1.541	7.737±2.121
Ranking Loss ↓	**0.124±0.045**	0.169±0.055	N/A	0.167±0.072	0.381±0.077	0.137±0.047	0.203±0.060
Average Precision ↑	**0.587±0.120**	0.518±0.135	0.479±0.143	0.533±0.139	0.368±0.091	0.564±0.131	0.568±0.124

的规模。本文采用随机交替下降法是因为 Hsieh 等人[12]指出实际运用中随机交替下降优于顺序交替下降。算法 1 给出算法流程。

算法 1 QPSolver($K, L, B, \gamma_a, \gamma_s, \mu$)

1: 输入 $K, L, B, \gamma_a, \gamma_s, \mu$
2: 令$\boldsymbol{\alpha}^0 = (\boldsymbol{\alpha}_1', \cdots, \boldsymbol{\alpha}_T')' = \mathbf{0}$, $\boldsymbol{\xi}^0 = \mathbf{0}$, $l = 1$
3: Repeat
4: 生成 1 到 T 的随机排列 s
5: For r = 1:T
6: t = s(r); 固定$\{\boldsymbol{\alpha}_q, \boldsymbol{\xi}_q\} \forall q \neq t$
7: 根据式(7)求解$\{\boldsymbol{\alpha}_t, \boldsymbol{\xi}_t\}$
8: end for
9: $\boldsymbol{\alpha}^l = (\boldsymbol{\alpha}_1', \cdots, \boldsymbol{\alpha}_T')'$, $\boldsymbol{\xi}^l = (\boldsymbol{\xi}_1', \cdots, \boldsymbol{\xi}_T')', l = l + 1$
10: 直到收敛，输出$\boldsymbol{\alpha}^l$，$\boldsymbol{\xi}^l$

4.2 固定$\boldsymbol{\alpha}, \boldsymbol{\xi}$求解$B$

给定$\boldsymbol{\alpha}, \boldsymbol{\xi}$，$B$的学习等价于求解如下优化问题，

$$\min_{B \in \mathcal{A}} \ \langle \hat{\boldsymbol{f}}' L \hat{\boldsymbol{f}}, B^{-1} \rangle \tag{8}$$

考虑到可行域$\mathcal{A}$的结构，根据文献[1]中的证明，式(8)的最优解是个闭式解(closed form)，可由下式给出，

$$B = \frac{\hat{\boldsymbol{f}}' L \hat{\boldsymbol{f}}}{trace(\hat{\boldsymbol{f}}' L \hat{\boldsymbol{f}})} \tag{9}$$

算法 2 给出 MASS 方法的整个算法流程。

算法 2 MASS 方法

1: 输入$K, L, B, \gamma_a, \gamma_s, \mu$
2: Repeat
3: 求解$\{\boldsymbol{\alpha}, \boldsymbol{\xi}\}$= QPSolver($K, L, B, \gamma_a, \gamma_s, \mu$)
4: 根据式(9)求解B
5: 直到收敛

5 实验结果

本文 MASS 方法将在两个真实数据上与多标记学习的多种典型方法进行实验比较，其中包括转化为多个两类问题的方法 BSVM[14]；转化为标记排序问题的方法 RANK-SVM [9]；考虑标记相关信息的方法 MLKNN[23]、BOOSTTEXTER[19]、ADTBOOST.MH[8]以及直推式方法 CNMF [15]。

实验参数设置如下：BOOSTTEXTER 和 ADTBOOST.MH 的轮数分别设置 500 和 50。当轮数继续增大时，这两类方法的性能没有发生太大的变化；RANK-SVM、MLKNN 以及 CNMF 方法均采用文献[9,23,15]中公布的最好参数；BSVM 的参数由交叉验证确定。MASS 方法的参数γ_a, γ_s, μ由交叉验证确定，矩阵B初始化为单位矩阵。

实验采用五种常用的多标记学习评价指标对算法性能进行评估：Hamming Loss、Ranking Loss、

表 4 基因功能分析数据 15%标记训练样本的结果

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	**0.206±0.003**	0.246±0.006	0.263±0.006	0.209±0.003	N/A	0.216±0.004	0.227±0.005
One-Error ↓	0.251±0.008	0.317±0.022	0.347±0.020	**0.245±0.005**	0.671±0.022	0.264±0.019	0.244±0.020
Coverage ↓	**6.599±0.113**	7.120±0.092	7.344±0.109	6.632±0.156	10.777±0.061	6.836±0.119	7.298±0.217
Ranking Loss ↓	**0.184±0.005**	0.222±0.007	N/A	**0.184±0.004**	0.476±0.006	0.196±0.004	0.211±0.010
Average Precision ↑	**0.744±0.005**	0.698±0.009	0.674±0.009	**0.744±0.004**	0.431±0.007	0.729±0.005	0.728±0.013

表 5 基因功能分析数据 10%标记训练样本的结果

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	**0.211±0.004**	0.251±0.005	0.275±0.009	0.215±0.003	N/A	0.221±0.005	0.231±0.004
One-Error ↓	0.255±0.012	0.325±0.024	0.357±0.024	**0.247±0.004**	0.722±0.017	0.264±0.011	0.255±0.016
Coverage ↓	6.684±0.141	7.262±0.146	7.568±0.168	**6.633±0.141**	10.842±0.073	6.953±0.148	7.332±0.194
Ranking Loss ↓	0.189±0.006	0.228±0.009	N/A	**0.188±0.005**	0.493±0.006	0.202±0.005	0.218±0.009
Average Precision ↑	**0.738±0.007**	0.691±0.011	0.663±0.011	0.737±0.005	0.418±0.005	0.722±0.005	0.719±0.010

表 6 基因功能分析数据 5%标记训练样本的结果

评价指标	MASS	BOOSTTEXTER	ADTBOOST.MH	RANK-SVM	CNMF	MLKNN	BSVM
Hamming Loss ↓	**0.218±0.004**	0.258±0.007	0.283±0.009	0.231±0.020	N/A	0.234±0.008	0.234±0.005
One-Error ↓	0.265±0.018	0.345±0.026	0.388±0.027	**0.252±0.003**	0.725±0.013	0.277±0.022	0.257±0.018
Coverage ↓	6.807±0.157	7.470±0.153	7.872±0.156	**6.759±0.211**	10.855±0.063	7.255±0.271	7.522±0.219
Ranking Loss ↓	**0.200±0.006**	0.242±0.011	N/A	**0.200±0.007**	0.501±0.006	0.218±0.010	0.224±0.010
Average Precision ↑	**0.725±0.007**	0.676±0.012	0.645±0.012	**0.725±0.005**	0.417±0.004	0.704±0.010	0.711±0.009

One-Error、Coverage、Average Precision。文献[19]中给出了这些评价指标的详细介绍。CNMF 方法无法提供 Hamming Loss 的结果，ADTBOOST.MH 方法的实现无法提供 Ranking Loss 的结果[23]。

实验将考察不同数量标记样本对性能的影响。实验分别抽取 5%、10%、15%的训练样本作为标记样本，余下为未标记样本。重复 30 次以上实验，汇报其平均结果。由于 CNMF 方法是直推式方法，不能直接对除测试样本以外的未见样本进行预测，实验中将最终测试样本作为 CNMF 训练时的未标记样本。

5.1 网页分类数据

第一个多标记学习任务是文本分类中的网页分类数据集[21,13,24]。网页数据从 yahoo.com 域名下收集得到，根据 yahoo 的高级类别分成 11 个数据集，每个网页隶属于多个 yahoo 的二级类别。每个数据集包括 2000 个训练样例和 3000 个测试样例，其中约 20%~45%的样例是多标记的。有关这 11 个数据集的详细介绍可参见文献[23]。

表 1-3 分别给出本文 MASS 方法与比较方法在 15%、10%、5%标记样本时的各个评价指标上的结果。加粗部分表示每个指标上的最佳性能，及与最佳性能在 95%置信区间 pairwise ttest 统计检验不显著差的性能。可以看到，除了在 Hamming Loss 指标上不如 BSVM 外，MASS 方法在不同数量标记样本的其它指标上都取得最好的性能。CNMF 没有取得良好的性能，可能原因是它的基本假设，即样本之间的相似性可由样本标记之间的相似性表出，在网页分类数据上不满足。

5.2 基因功能分析数据

第二个多标记任务是基因功能分析数据 Yeast。学习任务是预测酵母(Yeast Saccharomyces cerevisiae)的基因功能类[9]。其中每个基因通过微阵列基因表达数据(micro-array expression data)和系统发育谱(phylogenetic profile)联合表达，基因平均隶属于 4.24 个功能类。整个 Yeast 数据集共有 2417 个基因，每个基因用 103 维向量表示，共有 14 个不同标记。实验中随机抽取 500 个基因作为测试样例，余下的为训练样本，重复 30 次实验，汇报其平均结果。

从表 4-6 可以看出，MASS 与 RANK-SVM 取得最好的性能。具体而言，RANK-SVM 在 15%、10%、5%标记样本的 One-Error、10%、5%标记样本的 Coverage 及 10%标记样本的 Ranking-Loss 指标上取得最佳性能，而 MASS 在 15%、10%、5%标记样本的 Hamming-Loss、15%标记样本的 Coverage 及 10%标记样本的 Average Precision 指标上取得最佳性能。CNMF 仍然没有取得良好的性能。

6 结束语

本文研究了半监督多标记学习问题。这类学习问题广泛存在于多标记学习的实际应用中，直接利用现有多标记方法不能得到很好解决。本文提出了一种基于正则化的半监督多标记 MASS 方法，融合了丰富的未标记样本与标记关系信息来帮助提高性能，并给出快速优化算法。实验结果验证了 MASS 方法的有效性。基于本文的想法来进行半监督设置下的多示例多标记学习[24]将是一个值得研究的内容。

参 考 文 献

[1] Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning. Advances in neural information processing systems 19, 2007, 41-48

[2] Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. Bioinformatics 2006 22(7):830-836.

[3] Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research, 2006 7, 2399-2434.

[4] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. Pattern Recognition 2004, 37(9): 1757–1771

[5] Brinker, K., F ürnkranz, J., and Hullermeier, E. A unified model for multilabel classification and ranking. In Proceeding of the 17th European Conference on Artificial Intelligence, 2006, 489-493.

[6] Brinker, K., and Hullermeier, E. Case-based multilabel ranking. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007.702–707.

[7] Chen, G., Song, Y., Wang, F., and Zhang, C. Semi-supervised multi-label learning by solving a sylvester equation. SIAM International Conference on Data Mining, 2008, 410-419.

[8] Comit é, F. D., Gilleron, R., and Tommasi, M. Learning multi-label alternating decision tree from texts and data. In Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition, 2003, 35–49.

[9] Elisseeff, A., and Weston, J. A kernel method for multi-labelled classification. Advances in Neural Information Processing Systems 14. 2002, 681-687.

[10] Evgeniou, T., Micchelli, C., and Pontil, M. Learning Multiple Tasks with Kernel Methods. The Journal of Machine Learning Research, 2005, 6, 615-637.

[11] Ghamrawi, N., and McCallum, A. Collective multilabel classfication. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005.195–200.

[12] Hsieh, C., Chang, K., Lin, C., Keerthi, S., and Sundararajan,S. A dual coordinate descent method for large-scale linear SVM. Proceedings of the 25th international conference on Machine learning,2008, 408-415.

[13] Kazawa, H., Izumitani, T., Taira, H., and Maeda, E. Maximal margin labeling for multi-topic text categorization. Advances in Neural Information Processing Systems 17. 2005, 649–656.

[14] Joachims, T. Text categorization with support vector machines: learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, 1998, 137–142.

[15] Liu, Y., Jin, R., and Yang, L. Semi-supervised multilabel learning by constrained non-negative matrix factorization. In Proceedings of 21st National Conference on Artificial Intelligence, 2006, 421–426.

[16] Luo, Z., and Tseng, P. On the convergence of the coordinate descent method for convex differentiable minimization. Journal of Optimization Theory and Applications, 1992, 72, 7-35.

[17] McCallum, A. Multi-label text classification with a mixture model trained by EM. In Working Notes of the AAAI'99 Workshop on Text Learning, 1999.

[18] SchÄolkopf, B., and Smola, A. J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, 2002, MIT Press.

[19] Schapire, R. E., and Singer, Y. Boostexter: a boosting-based system for text categorization. Machine Learning 2000, 39(2-3):135–168.

[20] Sun, Y.-Y., Zhang Y., and Zhou Z.-H. Multi-label learning with weak label. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010, 593-598.

[21] Ueda, N., and Saito, K. Parametric mixture models for multi-labeled text. Advances in Neural Information Processing Systems 15. 2003,721–728.

[22] Zhang, M.-L., and Zhou, Z.-H. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering, 2006 (18), 1338-1351.

[23] Zhang, M.-L., and Zhou, Z.-H. Ml-knn: a lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7):2038-2048.

[24] Zhou, Z.-H., and Zhang, M.-L. Multi-instance multilabel learning with application to scene classification. Advances in Neural Information Processing Systems 19, 2007, 1609-1616.

[25] 孔祥南, 黎铭, 姜远, 周志华. 一种针对弱标记的直推式多标记分类方法. 计算机研究与发展, 2010, 47(8): 1392-1399.

李宇峰，男，1983 年生，博士研究生，研究方向：机器学习、数据挖掘。联系电话：(025)83685926，手机：13814178353。

黄圣君，男，1988 年生，博士研究生，研究方向：机器学习、数据挖掘。

周志华，男，1973 年生，教育部长江学者特聘教授（博导），研究领域：人工智能、机器学习、数据挖掘、模式识别等