ISSN 1000-1239/CN 11-1777/TP
47(8): 1392-1399, 2010

一种针对弱标记的直推式多标记分类方法

孔祥南　黎　铭　姜　远　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)
(kongxn@lamda.nju.edu.cn)

A Transductive Multi-Label Classification Method for Weak Labeling

Kong Xiangnan, Li Ming, Jiang Yuan, and Zhou Zhihua
(*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093)

Abstract　Multi-label learning deals with the problems when each object can be assigned to multiple categories simultaneously, which is ubiquitous in many real world applications, such as text classification, image scene classification and bioinformatics, etc. In traditional multi-label learning methods, classifiers are usually required to utilize a large amount of fully labeled training data in order to obtain good performances for multi-label classifications. However, in many real world tasks, obtaining partially labeled (weak labeled) training data is often much easier and costs less efforts than obtaining a large amount of fully labeled training data. To alleviate the assumption of large amount fully labeled training data used by traditional multi-label learning methods, the authors propose a new multi-label learning method for weak labeling (TML-WL). By reweighting the error functions on positive and negative labels of weak labeled data, TML-WL method can effectively utilize the weak labeled training data to replenish the missing labels. TML-WL method can also use the weak labeled training data to improve the classification performances on unlabeled data. Empirical studies on the real-world application of image scene classification show that the proposed method can significantly improve the performance of multi-label learning when the training data are weak labeled.

Key words　machine learning; multi-label classification; weak labeling; scene classification; transductive learning

摘　要　多标记学习主要解决一个样本可以同时属于多个类别的问题,它广泛适用于图像场景分类、文本分类等任务.在传统的多标记学习中,分类器往往需要利用大量具有完整标记的训练样本才能获得较好的分类性能,然而,在很多现实应用中又往往只能获得少量标记不完整的训练样本.为了更好地利用这些弱标记训练样本,提出一种针对弱标记的直推式多标记分类方法,它可以通过标记误差加权来补全样本标记,同时也能更好地利用弱标记样本提高分类性能.实验结果表明,该方法在弱标记情况下的图像场景分类任务上具有较好的性能提高.

关键词　机器学习;多标记学习;弱标记;图像场景分类;直推式学习

中图法分类号　TP181;TP391.41

收稿日期:2009-05-11;**修回日期**:2009-10-12

基金项目:国家自然科学基金项目(60975043);江苏省自然科学基金项目(BK2008018)、江苏省333高层次人才培养工程基金项目

通信作者:姜远(jiangyuan@nju.edu.cn)

0 引 言

在很多真实世界的分类问题中,一个样本往往可以同时属于多个不同的类别.比如在图像场景分类任务中,一幅图像会因为包含了多个语义概念,而同时拥有"天空"、"海洋"、"日落"等多个标记;在文本分类任务中,一个文档可能由于其内容涉及的不同方面而同时拥有"新闻"、"娱乐"、"教育"等多个主题.如何对这种存在标记多义性的数据进行有效学习,使得尽可能准确全面地预测出未知样本所包含的全部标记,已成为近年来机器学习研究的热点之一.

多标记学习(multi-label learning)是一种针对标记多义性样本进行学习的重要技术.目前,多种多标记学习方法被相继提出,包括基于集成学习的BOOSTEXTER[1]、基于核方法的RANK-SVM[2]、基于 k 近邻方法的ML-kNN[3]、基于BP神经网络的BP-MLL[4]、基于多示例多标记框架的INSDIF[5]等.这些多标记学习方法被成功地应用在文本分类[6-9]、基因功能分析[2,4]、自然场景分类[10]、相关规则挖掘[11-12]等任务中.

以往对多标记学习的研究大多是在训练样本标记完整的设置下进行的,即每个样本中所应该具有的标记都出现.然而,在很多现实世界的应用中,为每个样本提供其对应的完整标记往往相当困难.例如:在图像标注任务中,一幅图像往往拥有大量的候选标记类别,想要完整标注训练集中的每一样本就意味着需要人工查看每一幅图像的所有候选类别并逐一标注,确保无一遗漏.当数据规模和类别数目较大时,要获得完整标注的训练样本将需要消耗巨大的人力物力才能保证.事实上,在很多真实的应用中,收集大量标记不完整的弱标记(weak labeled)样本,则相对较为容易.在此,一个弱标记样本可能仅包含其对应所有标记中的一小部分,或者根本没有被标出任何的标记.

然而,现有的多标记学习方法则不能很好地针对这种弱标记样本进行有效学习.如果直接使用这些弱标记样本进行学习,则在学习中隐含假设了样本不包含的标记一定不属于这个样本,从而给训练集引入大量的噪声.为保证学习质量,通常在学习之前需要删除这些弱标记的样本.这使得训练集规模急剧减小,从而影响最终的学习性能,同时,也造成了数据资源的浪费.如果能够有效利用这些弱标记样本进行学习,将大大提高多标记学习在实际问题中的应用效果.

事实上,这些弱标记样本在学习过程中具有双重性质.对于出现的标记而言,这些样本是有标记样本;对于未出现的标记而言,这些样本则是无标记样本.针对弱标记样本进行学习需要能够对这些无标记样本进行有效利用.直推式学习(transductive learning)[13-14]是利用未标记数据学习的主流技术之一[15].通过对有标记样本和无标记样本进行学习,以期对无标记样本的标记给出尽可能准确的预测.

本文提出了一种针对弱标记样本进行学习的直推式多标记学习方法TWL-WL.该方法可以有效地利用弱标记样本的标记信息来进行多标记学习.在正则化框架下,通过最小化定义在弱标记样本上的加权损失函数,并采用输出函数的局部平滑性作为正则化项,获得所有弱标记样本在各个类别上的最优预测函数值.TML-WL方法可以用来为弱标记样本补全标记,同时也可以对需要预测的未标记样本进行预测.在图像的自然场景分类的应用表明,TML-WL方法能够有效地利用弱标记样本进行学习.

1 TWL-WL 方法

假设一个多标记数据集中有 n 个弱标记样本 $\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$.其中 $\boldsymbol{x}_i\in\mathcal{X}$ 对应的弱标记为 $\boldsymbol{y}_i=[y_{i1},y_{i2},\cdots,y_{im}]^{\mathrm{T}}\in\{0,1\}^m$,$m$ 为类别数目.当用户为样本 $\boldsymbol{x}_i$ 标出第 z 类标记时,$y_{iz}=1$,对于其他未标出的类别 $y_{iz}=0$.所有样本对应的弱标记矩阵为 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n]^{\mathrm{T}}$.

首先定义TML-WL方法对样本标记的实值输出函数为 $f:\mathcal{X}\mapsto\mathcal{R}^m$.设函数 f 对样本 $\boldsymbol{x}_i$ 在第 z 类的输出值为 f_{iz},定义 $\boldsymbol{f}_i=[f_{i1},f_{i2},\cdots,f_{im}]^{\mathrm{T}}$,矩阵 $\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]^{\mathrm{T}}$.然后,TML-WL方法定义一种针对弱标记情况的直推式多标记学习的误差函数为

$$E(f)=E_1(f)+\mu E_2(f),$$

其中,$\mu\geqslant 0$,为正则化参数.具体来说,TML-WL方法的优化目标为

$$\arg\min_f \frac{1}{2}\sum_z^m\sum_i^n M_{iz}(y_{iz}-f_{iz})^2+\frac{1}{2}\mu\sum_{ij}^n W_{ij}\left\|\frac{\boldsymbol{f}_i}{\sqrt{d_i}}-\frac{\boldsymbol{f}_j}{\sqrt{d_j}}\right\|^2,$$

其中 $E_1(f)$ 为针对弱标记的加权误差函数.与传统的处理方式不同,本文采用对弱标记样本误差加权

的方法来缓解用户标记不完整的影响. 即对用户已经标出的标记完全信任,误差权重设置为 1;而对于用户未标出的类别,该样本虽被视为不包含此类别的标记,但在误差函数上实行加权,权值为 $\theta(0\leqslant\theta\leqslant1)$,从而减少弱标记带来的标记噪声对方法模型造成的影响. 具体来说,加权误差函数 $E_1(f)$ 的定义为:

$$E_1(f)=\frac{1}{2}\sum_{z}^{m}\sum_{i}^{n}M_{iz}(y_{iz}-f_{iz})^2.$$

这里 M_{iz} 表示样本 $\boldsymbol{x}_i$ 在类别 z 上的误差权重,

$$M_{iz}=\begin{cases}1, & y_{iz}=1,\\ \theta, & y_{iz}=0.\end{cases}$$

所有样本的误差权重用矩阵形式表示为 $\boldsymbol{M}=[M_{iz}]$. 通过最小化加权误差函数 $E_1(f)$ 可以保证函数 f 在用户标记出的类别上输出值尽可能接近真实标记,同时又能减少未被用户标出的标记被当成负标记对函数 f 造成的影响.

然后根据样本图上的平滑假设,即相似样本的标记也相似,本文定义了误差函数 $E_2(f)$. 这里受到基于图的半监督学习算法的启发[16],首先在样本上建立一个 k 近邻图,图上的点代表数据集中的样本,边代表样本之间的相似度. 边上的权重 w_{ij} 由如下公式计算得到,

$$w_{ij}=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right),$$

这里 σ 为宽度参数,本文中固定为样本间的平均距离. 于是误差函数

$$E_2(f)=\frac{1}{2}\sum_{ij}^{n}W_{ij}\left\|\frac{\boldsymbol{f}_i}{\sqrt{d_i}}-\frac{\boldsymbol{f}_j}{\sqrt{d_j}}\right\|^2=\operatorname{tr}(\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{\frac{1}{2}})\boldsymbol{F}),$$

其中,$d_i=\sum_{k}^{n}W_{ik}$,设矩阵 $\boldsymbol{D}=\operatorname{diag}([d_1,\cdots,d_n])$,$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{\frac{1}{2}}$. 通过最小化函数 $E_2(f)$,可以在样本上获得较为平滑的标记输出.

综合上面的两方面考虑,本文使用了针对于弱标记的直推式多标记学习方法 TML-WL 的误差函数为 $E(f)=E_1(f)+\mu E_2(f)$. 这里 $E(f)$ 的梯度可以用矩阵形式表示为

$$\frac{\partial E(f)}{\partial \boldsymbol{F}}=\boldsymbol{M}\odot[\boldsymbol{F}-\boldsymbol{Y}]+\mu\boldsymbol{L}\boldsymbol{F},$$

$\odot$ 表示矩阵的按位相乘运算,求解最优条件 $\frac{\partial E(f)}{\partial \boldsymbol{F}}=0$,可得到关于 $\boldsymbol{f}_{\cdot z}$ 的线性方程组,

$$(\hat{\boldsymbol{M}}_{\cdot z}+\mu\boldsymbol{L})\boldsymbol{f}_{\cdot z}=\boldsymbol{c}_z, \tag{1}$$

其中,$z=1,2,\cdots,m$,$\hat{\boldsymbol{M}}_{\cdot z}=\operatorname{diag}(\boldsymbol{M}_{\cdot z})$,$\boldsymbol{C}=\boldsymbol{M}\odot\boldsymbol{Y}$,且 $\boldsymbol{c}_z=\boldsymbol{C}_{\cdot z}=\boldsymbol{M}_{\cdot z}\odot\boldsymbol{Y}_{\cdot z}$. 本文通过求解线性方程组(1)的最小二乘解得到 TML-WL 方法目标函数的近似最优解. 算法 1 给出了 TML-WL 的方法描述.

算法 1. TML-WL 方法.

输入: Instances, X; Weak label matrix, $\boldsymbol{Y}$; Weak label weight, θ; Smoothness, μ;

输出: Label output matrix, $\boldsymbol{F}$

① Construct k nearest neighbor graph on instances, and calculate similarity matrix $\boldsymbol{W}$ by $w_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2)$, then $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{\frac{1}{2}}$;

② Initialize error weight matrix $\boldsymbol{M}$, if the i-th instance is labeled with z-th label by user, $M_{iz}=1$; otherwise $M_{iz}=\theta$;

③ Solve the linear System (1), and calculate the label outputs $\boldsymbol{f}_{\cdot z}$ by

For $z=1$ to m

$(\hat{\boldsymbol{M}}_{\cdot z}+\mu\boldsymbol{L})\boldsymbol{f}_{\cdot z}=\boldsymbol{c}_z$

End

$\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]^{\mathrm{T}}$

2 实验测试

2.1 实验设置与评价准则

本文选取了文献[10]中所使用的多标记自然场景图像分类数据集,包含 2 400 幅图像. 图像特征首先将每幅图像转换到 CIE Luv 颜色空间,然后将图片分成 7×7 等宽的格子,分别计算各个颜色的一阶二阶距,得到颜色纹理特征共 294 维. 图像分别被标注为"beach","mountain","field","sunset"等 6 种场景类别.

为了衡量不同算法的多标记学习性能,本文采用了如下评价标准:

1. *RankingLoss*. 衡量当对样本的标记实值输出排序时,样本不具有的标记排在其真实标记前的错误情况所占比例.

$$RankingLoss(f,\mathcal{U})=\frac{1}{\mathcal{U}}\sum_{i\in\mathcal{U}}\frac{1}{|\boldsymbol{y}_i||\bar{\boldsymbol{y}}_i|}|\{y_1,y_2\}\in\boldsymbol{y}_i\times\bar{\boldsymbol{y}}_i\mid f(\boldsymbol{x}_i,y_1)\leqslant f(\boldsymbol{x}_i,y_2)\}|.$$

这里的 $\bar{\boldsymbol{y}}$ 表示真实标记集合 $\boldsymbol{y}$ 在所有类别集合 $\mathcal{Y}$ 中的补集,$RankingLoss\in[0,1]$,值越小则说明分类器对标记的排序性能越好.

2. *One-Error*. 衡量当对样本的标记实值输出排序时,排在最前面的标记不属于样本真实标记集合的比例.

$$One\text{-}Error(f,\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} Id(\operatorname{argmax}_{1\leqslant y\leqslant m} f(\boldsymbol{x}_i,y) \notin \boldsymbol{y}_i).$$

这里 $Id(\pi)=1$,如果条件 π 成立;否则 $Id(\pi)=0$. $One\text{-}Error\in[0,1]$,值越小则分类器对标记的排序性能越好.

3. *Coverage*. 衡量当对样本的标记实值输出排序时,排在最前面的多少个标记可以包含样本所有真实标记. *Coverage* 值越小,则分类器对标记的排序性能越好.

$$Coverage(f,\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \max_{y\in\boldsymbol{y}_i} r_f(\boldsymbol{x}_i,y) - 1.$$

4. *AveragePrecision*. 衡量当对样本的标记实值输出排序时,排在样本任一标记前面的标记中属于其真实标记集合的比例.

$$AveragePrecision(f,\mathcal{U}) = \frac{1}{\mathcal{U}}\sum_{i\in\mathcal{U}}\frac{1}{|\boldsymbol{y}_i|}\sum_{y\in\boldsymbol{y}_i}\frac{|\{y'\in\boldsymbol{y}_i \mid r_f(\boldsymbol{x}_i,y')\leqslant r_f(\boldsymbol{x}_i,y)\}|}{r_f(\boldsymbol{x}_i,y)}$$

$AveragePrecision\in[0,1]$,值越大则分类器对标记的排序性能越好.

以上评价指标分别从不同方面评价分类器的多标记分类性能. 为了方便分析结果,这里将 *AveragePrecision* 评价准则替换成 $1-AveragePrecision$,从而使得各个评价准则的性质统一,数值越低表示算法的性能越好.

本文使用了基于 k 近邻的多标记算法 ML-kNN 与 TML-WL 方法进行对比,以考察算法利用弱标记数据进行多标记学习的性能. 实验中,ML-kNN 算法使用文献[3]中报告的最优参数配置. TML-WL 方法的参数通过在单独的验证集上的 *RankingLoss* 性能进行选择,参数 μ 和 θ 的选择范围分别为{0.01,0.1,…,100}和{0.2,0.4,0.6,0.8,1.0}. 然后使用验证集上的最优参数在数据集上进行实验.

2.2 实验结果

实验中主要考虑多标记学习方法在弱标记条件下的两方面性能:一种是对弱标记样本补全标记的性能,另一种是利用弱标记训练样本来对未标记样本分类的性能.

首先,针对算法为弱标记数据补充完整标记的性能,本文在自然场景图像数据集中模拟了用户为图片提供弱标记的情况,将全部数据作为弱标记样本进行训练和测试. 这里值得注意的是,为了模拟弱标记情况,本文根据一定的比例(弱标记率)随机选出样本的一部分真实标记作为用户提供的弱标记. 这里的弱标记率(weak label rate)即为用户标出的标记与样本真实的完整标记集合大小的比例值. 例如当一个样本的完整的标记集合拥有 5 个真实标

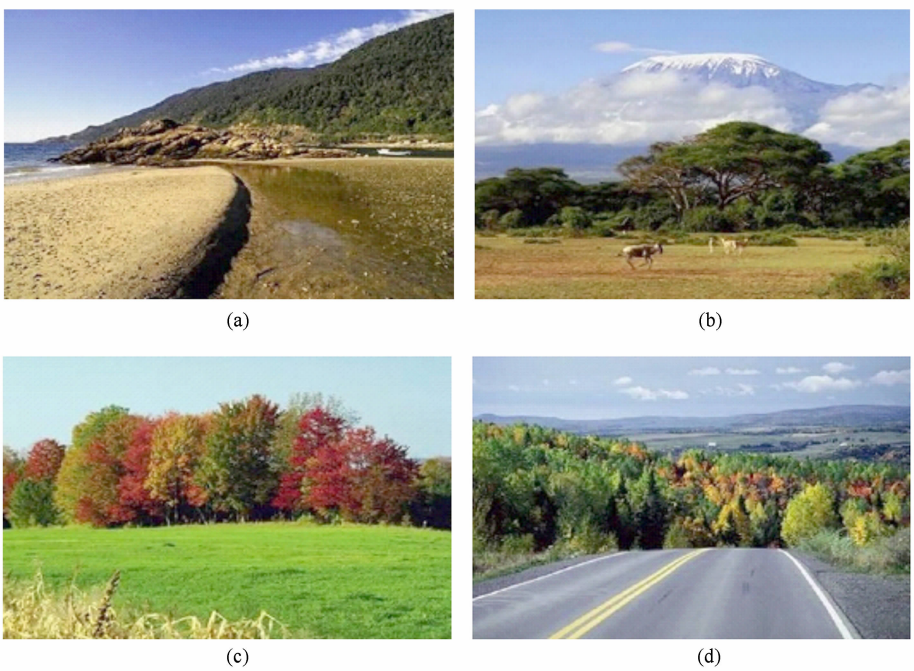

Fig. 1 The scene image sample with weak labels. (a) "Beach"; (b) "Mountain"; (c) "Field"; and (d) "Mountain".

图 1 自然场景分类图像的弱标记样例. (a)"海滩";(b) "山峰";(c)"绿地";(d) "山峰"

记,且弱标记率为 20%时,用户只为该样本随机标出其中的 1 个真实的标记. 图 1 中给出了自然场景图像的弱标记样例. 实验中我们测试了弱标记比例为 10%～90%的 9 种情况. 然后使用样本的完整真实标记来评估算法的性能.

图 2 中给出了两种方法在不同的弱标记率下 20 次随机实验的结果,记录了各项指标的平均值和标准差. 可以看出在为弱标记样本补全标记的任务上,TML-WL 取得了显著优于 ML-kNN 的分类性能.

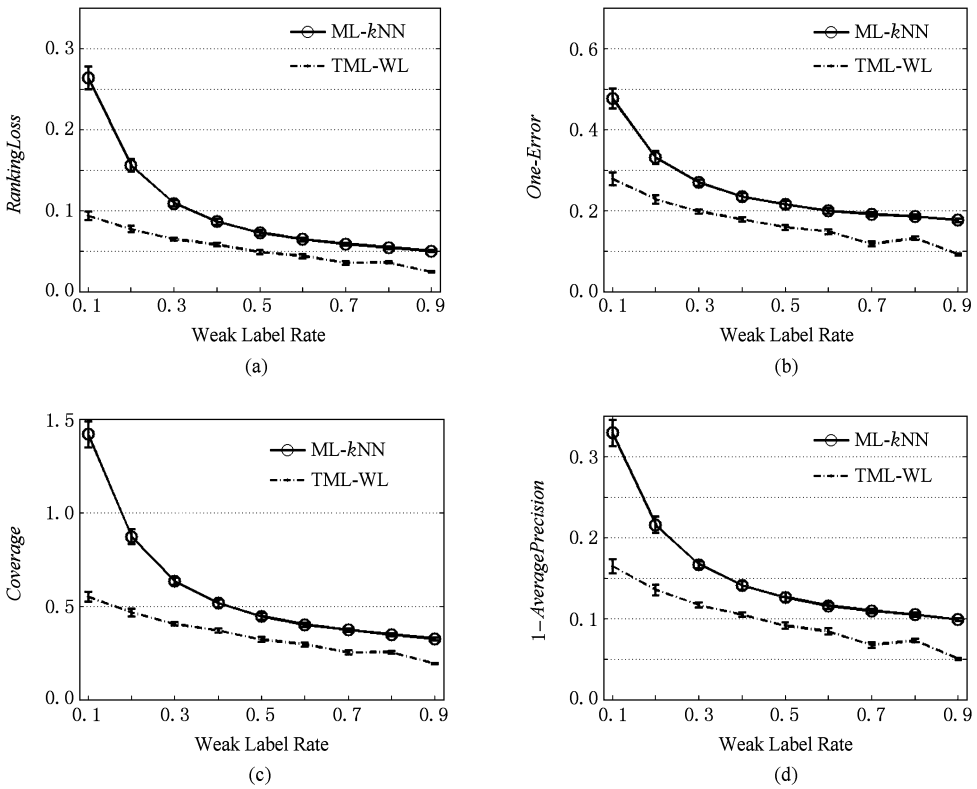

Fig. 2 The performance of replenishing labels for weak labeled data. (a) *RankingLoss*; (b) *One-Error*; (c) *Coverage*; and (d) 1−*Averageprecision*.

图 2 算法对弱标记样本补全标记的性能

然后,为了进一步评价算法利用弱标记训练样本来对未标记样本进行分类的性能,本文采用了第 2 种实验设置:在数据集上进行 5 倍交叉验证实验. 同样,为了模拟弱标记样本,本文使用了不同的弱标记率在训练集中模拟生成弱标记. 并在测试集评价算法性能时使用测试样本的完整真实标记来进行性能评估.

图 3 中给出了两种方法在不同的弱标记比例下的 5 倍交叉验证实验结果,记录了各项指标的平均值和标准差. 可以看出在自然场景图像分类任务上 TML-WL 取得了显著优于 ML-kNN 的平均性能. 在弱标记比例较高或者接近完整标注时(weak label rate 为 0.9),两种算法表现接近,TML-WL 在性能的各项均值上略优于 ML-kNN. 随着弱标记比例的减少,两种算法的性能虽然都会降低,但由于 TML-WL 考虑弱标记的影响,性能并未受到的较大改变. 可以看出,在用户仅仅标记出所有真实标记的 10%～20%时,TML-WL 受到弱标记的影响远远小于 ML-kNN 方法,且性能相对较好.

2.3 参数的影响

TML-WL 方法中参数 μ 是需要设置的正则化系数,用于平衡模型中两种误差的重要性. 为了测试参数 μ 对 TML-WL 方法的分类性能的影响,本节比较了不同的参数设置下得到的分类效果. 因在不同的弱标记率下算法的性能有所差异,本节实验中统一采用弱标记率为 20% 的实验设置进行测试,检验参数 μ 在 $\{0.01, 0.1, \cdots, 100\}$ 5 种情况下 TML-WL 方法的分类性能. 如图 4 所示,在所有的评价准则下,参数 μ 的最佳配置出现在 1～10 附近,此时 TML-WL 方法中的两种误差函数 $E_1(f)$ 和 $E_2(f)$ 的重要性接近. 当参数 μ 越大越接近 100 时,算法中的平滑性误差 $E_2(f)$ 占有了更大的权重,起到了主导作用,降低了弱标记信息的作用,从而性能严重下降. 同理,当参数 μ 越来越小,小于 10^{-2} 时,算法中对于标记产生的误差 $E_1(f)$ 的权重加大,$E_1(f)$ 渐

渐占据到主导作用.但由于其忽略了函数平滑性,势必导致函数的过拟合,因此性能也受到了影响.

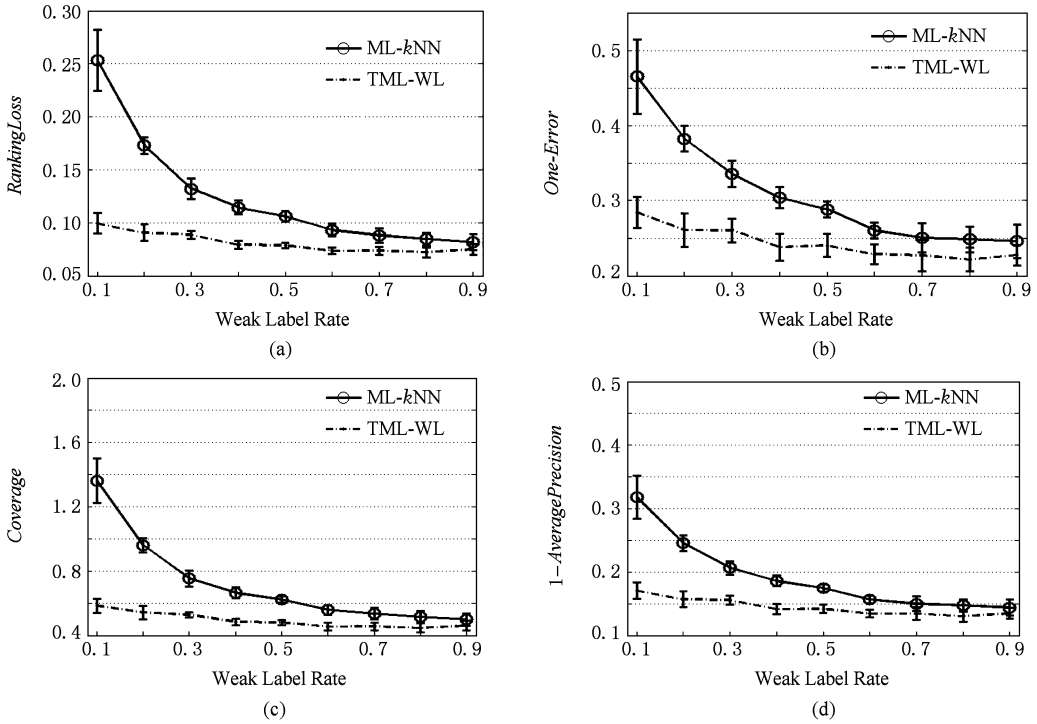

Fig. 3 The performance on unlabeled test data using weak labeled training data. (a) *RankingLoss*; (b) *One-Error*; (c) *Coverage*; and (d) 1−*AveragePrecision*.

图 3 算法利用弱标记数据预测未标记样本的性能.

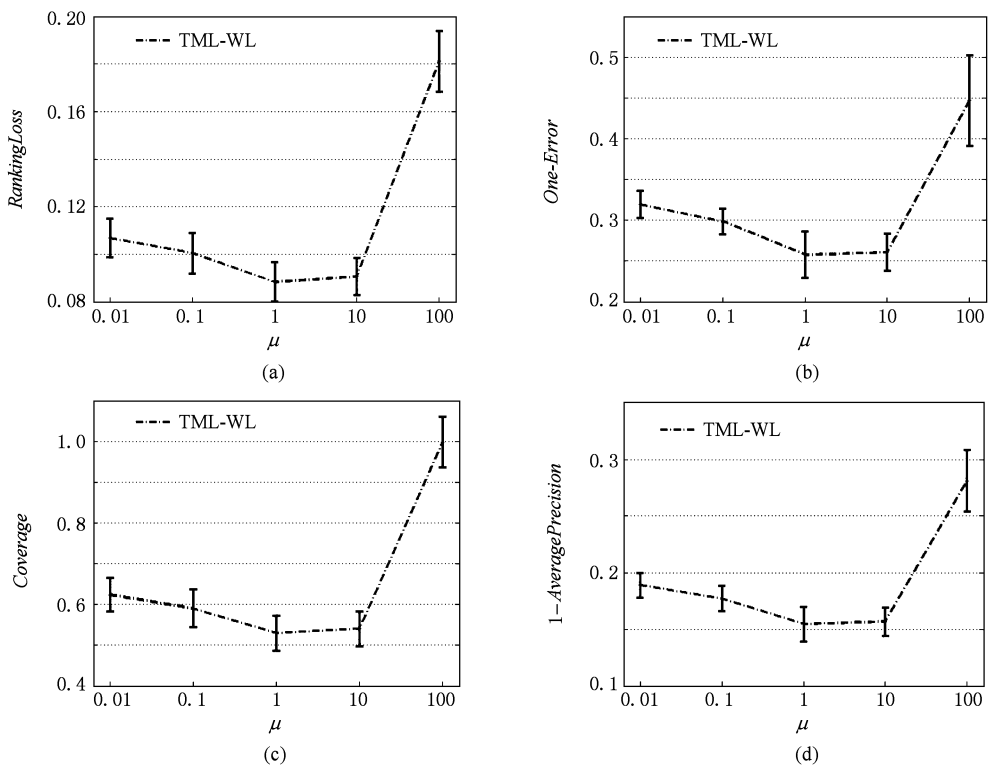

Fig. 4 The performance of TML-WL with different's. (a) *RankingLoss*; (b) *One-Error*; (c) *Coverage*; and (d) 1−*AveragePrecision*.

图 4 TML-WL 在不同 U 参数设置下的性能

3 结束语

本文针对多标记学习任务中仅能获得数据的少数不完整标记的情况,在正则化框架下提出了一种针对弱标记的直推式多标记分类方法 TML-WL. 该方法通过最小化定义在弱标记样本上的加权损失函数来更好地为弱标记样本补全标记. 在自然场景图像分类任务中,通过对比实验结果表明,该方法在弱标记情况下,受到标记噪声的影响较小. 相对传统的多标记方法而言,只需要更少的用户标注便可获得相对较好的多标记分类性能. 此外,一些以往研究工作也表明,在多标记学习中如何利用类别之间的相关性来进一步提高分类性能是一个更加困难的问题. 如何在弱标记情况下更好地利用类别相关信息来改进分类性能,仍将是以后需要进一步研究的问题.

参 考 文 献

[1] Schapire R E, Singer Y. Boostexter: A boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168

[2] Elisseeff A, Weston J. A kernel method for multi-labelled classification [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 681-687

[3] Zhang M -L, Zhou Z -H. Ml-*k*NN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048

[4] Zhang M -L, Zhou Z -H. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(10): 1338-1351

[5] Zhou Zhihua, Zhang Minling, Huang Shengjun, et al. MIML: A framework for learning with ambiguous objects, 0808.3231 [R]. Nanjing: National Key Laboratory for Novel Software Technology of Nanjing University, 2008 (in Chinese)
(周志华, 张敏灵, 黄圣君, 等. MIML: 一种从歧义对象中学习的框架, 0808.3231 [R]. 南京: 南京大学软件新技术国家重点实验室, 2008)

[6] Comite F D, Gilleron R, Tommasi M. Learning multi-label alternating decision tree from texts and data [C] //Proc of the 3rd Int Conf on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer, 2003: 35-49

[7] Gao S, Wu W, Lee C -H, et al. A MFoM learning approach to robust multiclass multi-label text categorization [C] //Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 329-336

[8] Kazawa H, Izumitani T, Taira H, et al. Maximal margin labeling for multi-topic text categorization [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 649-656

[9] McCallum A. Multi-label text classification with a mixture model trained by EM [C] //Working Notes of the AAAI'99 Workshop on Text Learning. Menlo Park, CA: AAAI, 1999: 1-7

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

[11] Rak R, Kurgan L, Reformat M. Multi-label associative classification of medical documents from MEDLINE [C] // Proc of the 4th Int Conf on Machine Learning and Applications. Washington, DC: IEEE Computer Society, 2005: 177-186

[12] Thabtah F A, Cowling P I, Peng Y. MMAC: A new multi-class, multi-label associative classification approach [C] // Proc of the 4th Int Conf on Data Mining. Washington, DC: IEEE Computer Society, 2004: 217-224

[13] Vapnik V N. Statistical Learning Theory [M]. New York: Wiley, 1998

[14] Jiang Yuan, She Qiaoqiao, Li Ming, et al. A transductive multi-label text categorization approach [J]. Journal of Computer Research and Development, 2008, 45(11): 1817-1823 (in Chinese)
(姜远, 佘俏俏, 黎铭, 等. 一种直推式多标记文档分类方法 [J]. 计算机研究与发展, 2008, 45(11): 1817-1823)

[15] Li Ming, Zhou Zhihua. Online semi-supervised learning with multi-kernel ensemble [J]. Journal of Computer Research and Development, 2008, 45(12): 2060-2068 (in Chinese)
(黎铭, 周志华. 基于多核集成的在线半监督学习方法[J]. 计算机研究与发展, 2008, 45(12): 2060-2068)

[16] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2004: 321-328

Kong Xiangnan, born in 1982. Master. His main research interests include machine learning and data mining.
孔祥南,1982 年生,硕士,主要研究方向为机器学习、数据挖掘等.

Li Ming, born in 1980. PhD. Received his B. Sc. degree in computer science from Nanjing University in 2003. His main research interests include machine learning and data mining, especially in learning with labeled and unlabeled examples.
黎 铭,1980 年生,博士,主要研究方向为机器学习、数据挖掘等.

Jiang Yuan, born in 1976. Received his PhD degree in computer science from Nanjing University in 2004. She is an associate professor at the Department of Computer Science & Technology, Nanjing University. Her main research interests include machine learning, information retrieval and data mining.

姜　远,1976 年生,博士,副教授,主要研究方向为机器学习、信息检索、数据挖掘等.

Zhou Zhihua, born in 1973. PhD, professor and PhD supervisor at the Department of Computer Science & Technology, Nanjing University. He is a senior member of China Computer Federation. His main research interests include artificial intelligence, machine learning, data mining, information retrieval, pattern recognition, evolutionary computation, and neural computation.

周志华,1973 年生,博士,教授,博士生导师,中国计算机学会高级会员,主要研究方向为人工智能、机器学习、数据挖掘、信息检索、模式识别、演化计算、神经计算等.

Research Background

In many real-world applications such as image scene classification, one instance usually belongs to multiple categories simultaneously, and therefore, it is important to classify an instance into a number of categories it belongs to by using multi-label learning technique. A large amount of fully labeled training examples are usually required for most of the existing multi-label learning approaches to achieve good performance. However, in most real applications only partially labeled (weak labeled) instances can be obtained, which always leads to unsatisfactory performance of the learned classifier if directly used. Thus, it is important to design a method that can automatically replenish the missing labels for the weak labeled data readily available for training to improve the performance of the classifiers. In this paper, we propose a transductive multi-label learning approach for weak labeling, which is able to exploit abundant weak labeled instances to help improve multi-label classification performance. Experimental results show that the proposed approach can improve the performance of multi-label learning when the training data are weak labeled. This work is supported by the National Natural Science Foundation of China (60635030), the National High Technology Research and Development Program of China (2007AA01Z169), the Jiangsu Science Foundation (BK2008018), and the Jiangsu 333 High-Level Talent Cultivation Program.