ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 77－87
DOI:10.16337/j.1004-9037.2015.01.007

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

偏标记学习研究综述

张敏灵[1,2]

(1.东南大学计算机科学与工程学院,南京,210096; 2.计算机网络和信息集成教育部重点实验室(东南大学),南京,210096)

摘　要:在弱监督信息条件下进行学习已成为机器学习领域的热点研究课题。偏标记学习作为一类重要的弱监督机器学习框架,适于多种实际应用问题的学习建模。在该框架下,每个对象在输入空间由单个示例(属性向量)进行刻画,而在输出空间与一组候选标记相关联,其中仅有一个为其真实标记。本文将对偏标记学习的研究现状进行综述,首先给出该学习框架的定义以及与相关学习框架的区别与联系,然后重点介绍几种典型的偏标记学习算法以及作者在该方面的初步工作,最后对偏标记学习进一步的研究方向进行简要讨论。

关键词:机器学习;弱监督信息;偏标记学习;候选标记;纠错输出编码

中图分类号:TP181　　**文献标志码**:A

Research on Partial Label Learning

Zhang Minling[1,2]

(1. School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China;
2. Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, 210096, China)

Abstract: In recent years, learning with weak supervision has become one of the hot research topics in machine learning. As one of the important weakly-supervised machine learning frameworks, partial label learning has been successfully applied to a number of real-world applications. In partial label learning, each object is described by a single instance (feature vector) in the input space. On the other hand, it is associated with a set of candidate labels among which only one is valid. The state-of-the-art on partial label learning researches is reviewed. Firstly, the problem definition on partial label learning as well as its differences and similarities with other related learning frameworks are given. Thenseveral representative partial label learning algorithms along with one of our recent progress on this topic are introduced. Finally, possible future investigations on partial label learning are briefly discussed.

Key words: machine learning; weakly-supervised information; partial label learning; candidate label; error-correcting output codes (ECOC)

基金项目:国家自然科学基金(61175049,61222309)资助项目;教育部新世纪优秀人才支持计划(NCET-13-0130)资助项目。
收稿日期:2014-12-10;**修订日期**:2015-01-03

引 言

在机器学习中,传统监督学习是研究得最多、应用最为广泛的一种学习框架[1]。假设 $\boldsymbol{x}=\mathbf{R}^d$ 为 d 维示例空间,$y=\{y_1,y_2,\cdots,y_q\}$为包含 q 个类别的标记空间,传统监督学习系统的任务是从训练集$\{(\boldsymbol{x}_i,y_i)|1\leqslant i\leqslant m\}$中学得函数 $f:x\rightarrow y$。其中,对于给定的训练样本$(\boldsymbol{x}_i,y_i)$,$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{id}]^{\mathrm{T}}\in x$ 代表一个 d 维的属性向量(示例),而 $y_i\in y$ 代表与 $\boldsymbol{x}_i$ 对应的类别标记。

基于上述框架,对于真实世界的每个对象,学习系统在输入空间基于一个示例刻画对象的性质,如文档对应的"词袋(bag-of-words)"向量。与此同时,在输出空间将示例与反映该对象语义信息的类别标记相关联,如文档隶属的主题。一般而言,类别标记作为监督信息蕴含了学习问题的语义和规律,是获得具有强泛化性能模型的关键因素。

传统监督学习框架在建模时采用强监督假设,即对象的类别标记信息是单一、明确的。对于满足上述假设的学习问题,传统监督学习框架已经取得了巨大的成功。值得注意的是,强监督假设虽然为学习建模的过程提供了便利,但却是对真实世界问题的一种简化处理,在许多情况下并不成立。实际上,受外部环境、问题特性以及物理资源等各方面因素的制约,学习系统往往只能从训练样本中获取有限的标记信息,即弱监督信息。如何在弱监督信息条件下有效地进行学习建模已成为机器学习领域的热点研究课题[2]。

偏标记学习是一类重要的弱监督机器学习框架[3],在该框架下,对象的类别标记不再具有单一性和明确性。目前,偏标记学习已在计算机视觉[4,5]、互联网[6]、生态信息学[7]等领域得到了成功应用。

1 框架定义

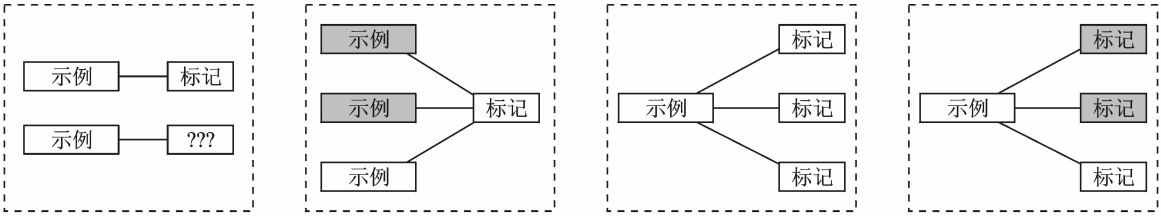
在偏标记学习框架下,每个对象可同时获得多个语义标记,但其中仅有一个标记反应了对象的真实语义,该形式的学习场景在现实世界问题中广泛存在。例如,在医疗诊断中,医生虽然可以排除病人患有某些疾病的可能,却往往难以从若干症状相似的疾病中给予确诊[8];在互联网应用中,用户可以自由地为各种在线对象提供标注,但对象获得的多个标注中可能仅有一个是正确的[9];再比如,人们可以从图像附属的标题文本中获取图像中各个人物名称作为语义标记,但对于图像中的特定人脸而言,它与各个语义标记(具体人物名称)的对应关系却并未给定[4],等等。

为了对上述形式的弱监督信息进行建模,研究者们提出了偏标记学习的概念[3]。采用与引言中相同的符号表示,偏标记学习框架的定义如下。

(1) 偏标记学习:假设 $x=\mathbf{R}^d$ 代表示例空间,$y=\{y_1,y_2,\cdots,y_q\}$代表(多类)标记空间。给定偏标记训练集 $D=\{(\boldsymbol{x}_i,S_i)|1\leqslant i\leqslant m\}$,其中 $\boldsymbol{x}_i\in x$ 为 d 维属性向量$[x_{i1},x_{i2},\cdots,x_{id}]^{\mathrm{T}}$,$S_i\subseteq y$ 为与 $\boldsymbol{x}_i$ 对应的候选标记集合,$\boldsymbol{x}_i$ 的真实标记 y_i 未知但满足条件 $y_i\in S_i$。基于此,偏标记学习系统的任务是基于训练集 D 学习得到多类分类器 $f:x\rightarrow y$。

根据弱监督信息的不同表现形式,图 1 将偏标记学习与 3 种主流的弱监督机器学习框架进行了对比,即半监督学习[10,11]、多标记学习[12,13]以及多示例学习[14,15]:

(2) 半监督学习:如图 1(a)所示,半监督学习是一种代表性的弱监督机器学习框架,此时训练集中仅有少量样本语义标记已知,而大量样本语义标记未知。因此,学习系统从训练集中可获取的监督信息十分有限。形式化地说,给定 L 个已标记训练样本 $L=\{(\boldsymbol{x}_i,y_i)|1\leqslant i\leqslant L\}$以及 U 个未标记训练样本 $u=\{\boldsymbol{x}_i|L+1\leqslant i\leqslant L+U\}$,其中 $\boldsymbol{x}_i\in x(1\leqslant i\leqslant L+U)$,$y_i\in y(1\leqslant i\leqslant L)$且 $L\ll U$,半监督学习系统的目标是从 $L\cup u$ 中学得函数 $f:x\rightarrow y$。



(a) 半监督学习 (b) 多示例学习 (c) 多标记学习 (d) 偏标记学习

(a) Semi-supervised learning (b) Multi-instance learning (c) Multi-label learning (d) Partial label learning

图 1 弱监督机器学习框架[3]

Fig. 1 Weakly-supervised learning framework

(3) 多示例学习:如图 1(b)所示,多示例学习是另一种代表性的弱监督机器学习框架,此时训练集中的每个样本采用示例包的表示形式,样本的语义标记定义在包层次而非示例层次,包为正包当且仅当包中含有正例。因此,学习系统从训练集中获取的监督信息较为有限,即正包中虽然包含正例但并未被明确标识,难以与包中的伪正例加以区分。形式化地说,给定多示例训练集 $D=\{(B_i,y_i)\mid 1\leqslant i\leqslant m\}$,其中 $B_i=\{x_i^j\mid 1\leqslant j\leqslant n_i, x_i^j\in x\}$ 为包含 n_i 个示例(属性向量)的包,$y_i\in y(y=\{-1,+1\})$ 为与包 B_i 对应的类别标记。基于此,多示例学习系统的任务是基于训练集 D 学习得到多示例分类器 $f:2^x\rightarrow y$。

(4) 多标记学习:如图 1(c)所示,在多标记学习框架下,每个对象可同时具有多个正确的语义标记,学习系统的目标是预测未见对象的标记集合。从表面上看,每个样本由于具有多个语义标记,与之对应的监督信息显得十分充分。然而,从形式化的角度看,如果将每种可能的标记集合看作一个类别,则多标记学习在本质上对应于一个多类学习问题,其输出空间的大小(即包含的类别数)具有指数规模。面对如此庞大的输出空间,学习系统从训练样本中获取的监督信息将显得十分有限,许多标记集合在训练集中仅对应于少量样本、甚至从未出现。形式化地说,给定多标记训练集 $D=\{(\boldsymbol{x}_i,Y_i)\mid 1\leqslant i\leqslant m\}$,其中 $\boldsymbol{x}_i\in x$ 为示例而 $Y_i\subseteq y$ 为与 $\boldsymbol{x}_i$ 对应的一组类别标记。基于此,多标记学习系统的任务是基于训练集 D 学习得到多标记分类器 $f:x\rightarrow 2^y$。

根据以上分析,偏标记学习与各主流弱监督学习框架的主要不同在于:(1)在半监督学习中,训练样本具有完全明确(单个真实标记)或者完全未知(无标记)的语义信息,而偏标记学习假设每个对象具有一个候选标记集合;(2)在多示例学习中,训练样本的标记信息明确但与示例的对应关系不明确,而偏标记学习假设每个对象具有单示例表示但与候选标记的对应关系不明确;(3)在多标记学习中,训练样本具有多个真实标记,且学习目标是获得从示例到标记集合的映射,而偏标记学习假设训练样本的真实标记包含于候选标记集合中,且学习目标是获得从示例到标记的映射。实际上,现实世界的一些应用问题采用偏标记学习框架进行描述显得更加自然[4,16-20]。

2 学习算法

在偏标记学习框架下,学习系统面临的监督信息不再具有单一性和明确性,真实的语义信息湮没于候选标记集合中,使得对象的学习建模变得十分困难。如 Cour 等人[3]近期报告的实验结果中,当候选标记集合较大时,学习系统在未见样本上的(多类)泛化精度甚至低于 30%。

为了设计有效的偏标记学习算法,一种直观的思路是对偏标记对象的候选标记集合进行消歧。采用该思路,现有的偏标记学习算法主要采用两种不同的消歧策略,即基于辨识的消歧以及基于平均的消歧。基于辨识的消歧将偏标记对象的真实标记作为隐变量,采用迭代的方式优化内嵌隐变量的目标函数实现消歧[7,9,21];基于平均的消歧赋予偏标记对象的各个候选标记相同的权重,通过综合学习模型在

各候选标记上的输出实现消歧[3,22]。

2.1 辨识消歧策略

2.1.1 极大似然估计方法[9]

如第1节所述,偏标记学习系统的任务是学得多类分类器 $f: x \to y$。假设分类器具有参数化的表达形式,即

$$f(x) = \arg\max_{y \in y} p(y \mid x, \boldsymbol{\theta}) \tag{1}$$

式中:$p(y|x,\boldsymbol{\theta})$代表样本 x 具有类别标记 y 的后验概率,$\boldsymbol{\theta}$ 为模型的参数向量。给定偏标记训练集 $D=\{(\boldsymbol{x}_i, S_i) \mid 1 \leqslant i \leqslant m\}$,其中 $\boldsymbol{x}_i \in x, S_i \subseteq y$。根据极大似然估计准则,当训练集中的样本满足条件独立性时,模型的最优参数可通过求解如下问题获得

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p(y \in S_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log\left(\sum_{y \in S_i} p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})\right) \end{aligned} \tag{2}$$

为了对似然函数(2)进行优化,一种常规做法是基于期望最大化算法[23]对模型参数进行迭代更新,直至获得(局部)最优解。

具体来说,假设在当前迭代轮次参数向量取值为 $\boldsymbol{\theta}^{(t)}$。在 E-step 中,EM 算法基于当前参数向量 $\boldsymbol{\theta}^{(t)}$ 对标记的后验概率分布进行估计

$$\hat{p}(y \mid \boldsymbol{x}_i) = \begin{cases} \dfrac{p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)})}{\sum\limits_{y' \in S_i} p(y' \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)})} & \forall y \in S_i \\ 0 & \text{其他} \end{cases} \tag{3}$$

值得注意的是,E-step 在一定程度上实现了对偏标记对象候选标记集合 S_i 的消歧操作,即:非候选标记 $y \notin S_i$ 的后验概率 $\hat{p}(y|\boldsymbol{x}_i)$ 为0,候选标记 $y \in S_i$ 的后验概率 $\hat{p}(y|\boldsymbol{x}_i)$ 体现了该标记成为样本真实标记的置信度。

如式(3)所示,E-step 估计所得的后验概率分布满足 $\sum\limits_{y \in S_i} \hat{p}(y|\boldsymbol{x}_i) = 1$。记式(2)中目标函数为 $L(\boldsymbol{\theta}) = \sum\limits_{i=1}^{m} \log\left(\sum\limits_{y \in S_i} p(y|\boldsymbol{x}_i, \boldsymbol{\theta})\right)$,根据 Jensen 不等式可得 $L(\boldsymbol{\theta})$ 的下界函数 $A(\boldsymbol{\theta})$

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \log\left(\sum_{y \in S_i} \hat{p}(y \mid \boldsymbol{x}_i) \cdot \frac{p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{\hat{p}(y \mid \boldsymbol{x}_i)}\right) \geqslant \\ &\sum_{i=1}^{m} \sum_{y \in S_i} \hat{p}(y \mid \boldsymbol{x}_i) \cdot \log \frac{p(y \mid \boldsymbol{x}_i, \theta)}{\hat{p}(y \mid \boldsymbol{x}_i)} = A(\boldsymbol{\theta}) \end{aligned} \tag{4}$$

结合式(4)与式(3),$L(\boldsymbol{\theta}) \geqslant A(\boldsymbol{\theta})$ 且当 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ 时该不等式取等号。基于此,在 M-step 中 EM 算法通过最大化下界函数 $A(\boldsymbol{\theta})$ 对参数向量进行更新

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) \tag{5}$$

根据式(4)与式(5),$L(\boldsymbol{\theta}^{(t+1)}) \geqslant A(\boldsymbol{\theta}^{(t+1)})$,$A(\boldsymbol{\theta}^{(t+1)}) \geqslant A(\boldsymbol{\theta}^{(t)})$ 且 $A(\boldsymbol{\theta}^{(t)}) = L(\boldsymbol{\theta}^{(t)})$。因此,基于每轮 EM 迭代,更新后的参数向量可以实现目标函数值的递增,即 $L(\boldsymbol{\theta}^{(t+1)}) \geqslant L(\boldsymbol{\theta}^{(t)})$。

算法1给出了基于极大似然估计的 PL-EM 算法[9]的伪码描述。给定偏标记训练集 D,算法首先对参数向量进行初始化(步骤1);然后,算法基于迭代的方式在 E-step 中对候选标记进行消歧(步骤4~6),并在 M-step 中更新参数向量(步骤7);最后,算法根据最大化后验概率准则预测测试样本的概念标记(步骤10~11)。

为了使用 PL-EM 算法,需要对后验概率模型 $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ 具体的参数化形式进行实例化,例如采用最

大熵(Maximum entropy, ME)模型[8,9,24]。进一步地,还可通过引入隐变量的方式采用概率图模型对后验概率模型进行刻画[7,25]。

算法 1 PL-EM 算法[9]的伪码描述

y^* =PL-EM (D, T, x^*)

1. Initialize parameter vector $\theta^{(0)}$;

2. $t=0$;

3. WHILE $t\leqslant T$ DO

4. FOR $i=1$ to m DO

5. Compute $\hat{p}(y|x_i)$ according to Eq. (3);

6. ENDFOR

7. Update $\theta^{(t+1)}$ according to Eq. (5);

8. $t=t+1$;

9. ENDWHILE

10. Let $\theta=\theta^{(T)}$;

11. Return $y^*=f(x^*)$ according to Eq. (1)。

2.1.2 最大化间隔方法[21]

给定标记空间 $y=\{y_1,y_2,\cdots,y_q\}$,设学习系统包含 q 个线性分类器$\{\boldsymbol{w}_j\mid \boldsymbol{w}_j\in\mathbf{R}^d, 1\leqslant j\leqslant q\}$,此时所需的多类分类器 f 对应于

$$f(x)=\arg\max_{y_j\in y}\boldsymbol{w}_j^{\mathrm{T}}\cdot\boldsymbol{x} \tag{6}$$

不失一般性,通过将每个示例 $\boldsymbol{x}$ 扩展一维取值恒为 1 的属性,等价于在各类别标记 y_j 的线性模型 w_j 中引入相应的偏置项。

最大化间隔是线性模型优化的常规策略,给定模型参数 $\boldsymbol{\omega}=[\boldsymbol{w}_1^{\mathrm{T}},\boldsymbol{w}_2^{\mathrm{T}},\cdots,\boldsymbol{w}_q^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{d\times q}$,其目标函数通常具有如下的表示形式

$$\min_{\boldsymbol{\omega}} L(\boldsymbol{\omega},D)+\lambda\cdot\Omega(\boldsymbol{\omega}) \tag{7}$$

式中:$L(\boldsymbol{\omega},D)$用于考察模型在训练样本上的经验损失,而 $\Omega(\boldsymbol{\omega})$用于考察模型的复杂度,参数用于平衡经验损失与模型复杂度对目标函数的影响。

给定偏标记训练集 D,需要选择合适的经验损失函数 $L(\boldsymbol{\omega},D)$以及模型复杂度函数 $\Omega(\boldsymbol{\omega})$以体现偏标记学习问题的特性。一般而言,模型复杂度可采用 L2-norm 正则函数:$\Omega(\boldsymbol{\omega})=\frac{1}{2}\|\boldsymbol{\omega}\|^2$。对于经验损失函数,该函数的选择需反映线性模型对每个偏标记训练样本$(\boldsymbol{x}_i,S_i)$的分类能力。

在偏标记学习问题中,对象的真实标记隐含于候选标记集合中。给定一组线性模型,一种直观确定真实标记的方法是将候选标记集合中线性输出值最大的标记作为真实标记。记$\overline{S}_i=y\backslash S_i$ 为非候选标记集合,则式(7)所示的优化问题可重写为

$$\begin{aligned}&\min_{\boldsymbol{\omega},\boldsymbol{\xi}}\frac{\lambda}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{m}\sum_{i=1}^{m}\xi_i\\ \text{s.t.}\quad&\max_{y_j\in S_i}\boldsymbol{w}_j^{\mathrm{T}}\cdot\boldsymbol{x}_i-\max_{y_k\in\overline{S}_i}\boldsymbol{w}_k^{\mathrm{T}}\cdot\boldsymbol{x}_i\geqslant 1-\xi_i\\ &\xi_i\geqslant 0\quad\forall i\in\{1,2,\cdots,m\}\end{aligned} \tag{8}$$

令 $\hat{y}_i$ 与$\overline{y}_i$ 分别代表 S_i 与$\overline{S}_i$ 中线性输出值最大的类别标记,即

$$\begin{aligned}\hat{y}_i&=\arg\max_{y_j\in S_i}\boldsymbol{w}_j^{\mathrm{T}}\cdot\boldsymbol{x}_i\\ \overline{y}_i&=\arg\max_{y_k\in\overline{S}_i}\boldsymbol{w}_j^{\mathrm{T}}\cdot\boldsymbol{x}_i\end{aligned} \tag{9}$$

式(8)中的第1项约束条件要求 $\hat{y}_i$ 的模型输出大于 $\bar{y}_i$(即所有非候选标记)的模型输出。相应地,松弛变量 $\boldsymbol{\xi}=\{\xi_1,\xi_2,\cdots,\xi_m\}$ 反映了各偏标记训练样本的分类间隔,其取值越小,则分类间隔越大。上述约束条件将 $\hat{y}_i$ 作为偏标记对象的真实标记,在一定程度上实现了对候选标记集合 S_i 的消歧操作。此外,当 $|S_i|=1$ 时偏标记对象退化为单标记对象,而该约束条件即为传统多类支持向量机[26]中的最大化间隔条件。对于如式(8)所示约束条件含算子的优化问题,可以基于次梯度技术如 Pegasos 方法[27]进行求解。Pegasos 方法通过迭代的方式实现模型更新,在迭代的每一轮中交替进行(次)梯度下降和投影操作。

为了便于后续的伪码描述,定义属性空间映射函数 $\Phi: x\times y\to\mathbf{R}^{d\times q}$。其中,给定"示例-标记"配对 $(x,y)\in x\times y$,该函数将其映射至新的属性向量 $\boldsymbol{\Phi}(x,y)$

$$\boldsymbol{\Phi}(x,y)=\begin{bmatrix} x\cdot[y=y_1] \\ x\cdot[y=y_2] \\ \vdots \\ x\cdot[y=y_q] \end{bmatrix} \tag{10}$$

式中:当谓词 π 成立时 $[\pi]$ 取值为1,否则取值为0。基于此,示例 x 在第 j 个线性模型 $\boldsymbol{w}_j$ 上的输出 $\boldsymbol{w}_j^{\mathrm{T}}\cdot\boldsymbol{x}_i$ 等价于 $\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{\Phi}(x,y_j)$。

算法2给出了基于最大化间隔策略的 PL-SVM 算法[21]的伪码描述。给定偏标记训练集 D,算法首先对线性模型进行初始化(步骤1);然后,算法基于迭代的方式依次进行(次)梯度下降操作(步骤4~6)以及投影操作(步骤7);最后,算法根据所得的线性模型预测测试样本的概念标记(步骤10)。

算法2 PL-SVM 算法[21]的伪码描述

$y^*=\text{PL-SVM}(D,\lambda,T,x^*)$

1. Initialize weight vectors $\boldsymbol{\omega}^{(1)}$, Such that $\|\boldsymbol{w}^{(1)}\|\leqslant 1/\sqrt{\lambda}$;
2. FOR $t=1$ to T DO
3. Let $\boldsymbol{\omega}=\boldsymbol{\omega}^{(t)}$;
4. Set $A=\{(x_i,S_i)\mid\xi_i>0,1\leqslant i\leqslant m\}$, where $\xi_i=1-(\max\limits_{y_j\in S_i}\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{\Phi}(x,y^j)-\max\limits_{y_k\in\bar{S}_i}\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{\Phi}(x,y^k))$;
5. Set $\eta_t=\dfrac{1}{\lambda t}$;
6. Set $\boldsymbol{\omega}^{(t+1)}=(1-\eta_t\lambda)\cdot\boldsymbol{\omega}+\dfrac{\eta_t}{m}\cdot gd$, where $gd=\sum\limits_{(x_i,S_i)\in A}\boldsymbol{\Phi}(x_i,\hat{y}_i)-\boldsymbol{\Phi}(x_i,\bar{y}_i)$ with $\hat{y}_i$ and $\bar{y}_i$ being identified by Eq. (9);
7. Set $\boldsymbol{\omega}^{(t+1)}=\min\left(1,\dfrac{1/\sqrt{\lambda}}{\|\boldsymbol{\omega}\|}\right)\cdot\boldsymbol{\omega}$;
8. ENDFOR
9. Let $\boldsymbol{\omega}=\boldsymbol{\omega}^{(T+1)}$;
10. Return $y^*=f(x^*)$ according to Eq. (6)。

2.2 平均消歧策略

2.2.1 k 近邻方法[22]

k 近邻是一种代表性的惰性学习算法[28],该算法无需任何模型假设,直接利用近邻样本估计测试样本的输出,在一定程度上避免错误的模型假设对学习带来的不利影响。给定偏标记训练集 D 以及测试样本 $\boldsymbol{x}^*\in x$,k 近邻算法首先计算 $\boldsymbol{x}^*$ 与各训练示例 $\boldsymbol{x}_i$ 之间的距离

$$d_i=\text{dist}(\boldsymbol{x}^*,\boldsymbol{x}_i) \tag{11}$$

式中距离函数 dist 可根据当前学习问题的特点进行选择。

将训练集中的样本根据 d_i 取值的升序进行排序。设$(\boldsymbol{x}_{i_r}, S_{i_r})$代表排在第 r 位的训练样本,则 $\boldsymbol{x}^*$ 在训练集中的 k 近邻可用如下集合表示

$$L=\{(x_{i_r}, S_{i_r}) \mid 1 \leqslant r \leqslant k\} \tag{12}$$

此时,可通过对近邻样本的候选标记集合进行加权投票以确定测试样本的类别标记

$$f(\boldsymbol{x}^*)=\arg\max_{y \in y} \sum_{r=1}^{k} \omega_r \cdot [y \in S_{i_r}] \tag{13}$$

其中,式(13)中的条件$[y \in S_{i_r}]$在一定程度上实现了对候选标记集合 S_{i_r} 的消歧操作,即:隶属于示例 x_{i_r} 的各候选标记 $y \in S_{i_r}$ 具有相同的投票权值 ω_r。设置投票权值 ω_r 的基本原则是近邻样本与 x^* 距离越小则权值越大,例如 $\omega_r=k-r+1$ 或 $\omega_r=1-d_r/(\sum_{r=1}^{k} d_r)(1 \leqslant r \leqslant k)$。

算法 3 给出了基于 k 近邻策略的 PL-kNN 算法[22]的伪码描述。给定偏标记训练集 D,算法首先计算测试样本与各训练示例之间的距离(步骤 1~3);然后,算法根据计算所得的距离确定其 k 近邻样本(步骤 4);最后,算法通过对近邻样本候选标记的加权投票预测测试样本的概念标记(步骤 5)。

算法 3 PL-kNN 算法[22]的伪码描述

$y^*=$PL-kNN(D, k, dist, x^*)

1. FOR $i=1$ to m DO
2. Compute distance d_i according to Eq. (11);
3. ENDFOR
4. Identify kNN set (i.e. L) according to Eq. (12);
5. Return $y^*=f(x^*)$ according to Eq. (13)。

2.2.2 *凸优化方法*[3]

不失一般性,设 $g_j: x \to \mathbf{R}$ 为与第 j 类对应的二类分类器,则多类分类器 f 可由上述二类分类器 $g_j(1 \leqslant j \leqslant q)$按如下方式获得

$$f(x)=\arg\max_{y_j \in y} g_j(x) \tag{14}$$

基于特定的二类凸损失函数 $\Psi: \mathbf{R} \to \mathbf{R}^+$,可在给定的偏标记训练集 D 上定义如下的经验损失目标函数

$$\begin{aligned} L_\Psi(D) &= \sum_{i=1}^{m} L_\Psi(\boldsymbol{x}_i, S_i) \\ &= \sum_{i=1}^{m} \left[\Psi\left(\frac{\sum_{y_j \in S_i} g_j(\boldsymbol{x}_i)}{|S_i|} \right) + \sum_{y_k \in \bar{S}_i} \Psi(-g_k(\boldsymbol{x}_i)) \right] \end{aligned} \tag{15}$$

式(15)第 1 个函数项 $\Psi(\sum_{y_j \in S_i} g_j(\boldsymbol{x}_i)/|S_i|)$在一定程度上实现了对候选标记集合 S_i 的消歧操作,即:偏标记样本在候选标记集合上的经验损失由各候选标记的模型输出的平均值决定。此外,考察学习系统如下的偏标记损失(partial 0/1 loss)

$$L_p(D)=\sum_{i=1}^{m} [\arg\max_{y_j \in y} g_j(\boldsymbol{x}_i) \notin S_i] \tag{16}$$

当凸损失函数 Ψ 为递减函数且 $\Psi(0) \geqslant 1$ 成立时,式(15)所示的经验损失 $L_\Psi(D)$ 与式(16)所示的偏标记损失 $L_p(D)$ 满足 $L_p(D) \leqslant \frac{1}{2} L_\Psi(D)$[3]。因此,当 Ψ 具有上述性质时可以将 $L_\Psi(D)$ 作为 $L_p(D)$ 的替代损失进行优化以获得最终的预测模型。满足上述性质的常用凸损失函数包括铰链损失 $\Phi(x)=\max(0, 1-x)$,指数损失 $\Phi(x)=e^{-x}$ 等。

记 Φ 为属性空间映射函数 $\Phi: x \times y \to R^{d \times q}$，其定义如式(10)所示。对于偏标记训练集 D，定义如下的正样本集 D^+ 以及负样本集 D^-

$$D^+ = \left[\frac{1}{|S_i|} \sum_{y_j \in S_i} \Phi(x_i, y_j) \mid 1 \leqslant i \leqslant m \right] \tag{17}$$

$$D^- = [\Phi(x_i, y_k) \mid 1 \leqslant i \leqslant m, y_k \in \overline{S}_i] \tag{18}$$

因此，D^+ 与 D^- 中包含的正/负样本数分别为 m 与 $\sum_{i=1}^{m}(q - |S_i|)$。设 B_Ψ 为优化某凸损失函数 Ψ 的二类学习算法(如优化铰链损失的 SVM)，则基于 B_Ψ 学习二类分类器 $h: x \to R(h \leftarrow B_\Psi(D^+ \cup D^-))$ 等价于优化式(15)所示的经验损失目标函数。

算法 4 给出了基于凸优化策略的偏标记凸损失(Convex loss for partial labels, CLPL)算法[3]的伪码描述。给定偏标记训练集 D，算法首先将其转化为一组正样本集与负样本集(步骤 1～2)；然后，算法通过优化特定的凸损失函数学得二类分类器(步骤 3～4)；最后，算法根据所得的凸优化模型预测测试样本的概念标记(步骤 5)。

算法 4 CLPL 算法[3]的伪码描述

$y^* = \mathrm{CLPL}(D, B_\Psi, \boldsymbol{x}^*)$

1. Derive the set of positive training examples D^+ from D^- according to Eq. (17);
2. Derive the set of negative training examples D^- from D according to Eq. (18);
3. Induce binary classifier $h \leftarrow B_\Psi(D^+ \cup D^-)$;
4. Set $g_j (1 \leqslant j \leqslant q)$ as $g_j(x) = h(\Phi(x, y_j))$;
5. Return $y^* = f(x^*)$ according to Eq. (14)。

2.3 非消歧策略

基于消歧策略可以实现对偏标记对象候选标记集合的利用，从而构建所需的偏标记学习系统。然而，由于对象的真实标记隐含于候选标记集合中，基于消歧的偏标记学习技术会受到“伪标记(即 S_i/y_j)”带来的不利影响。对于辨识消歧策略(2.1 节)，算法每一轮迭代过程中辨识出的类别标记可能为对象的伪标记而非其真实标记；相应地，对于平均消歧策略(2.2 节)，对象真实标记的模型输出可能湮没于其伪标记的模型输出中。

偏标记学习系统的最终目标是学习得到多类分类器 $f: x \to y$。在传统监督学习框架下，实现该目标最流行的机制之一是将多类学习问题分解为多个二类学习问题。其中，“一对多(one-vs-rest)”机制将多类学习问题分解为 q 个二类学习问题，每个二类学习问题将 $y_j (1 \leqslant j \leqslant q)$ 作为正类而将其他类作为负类；相应地，“一对一(one-vs-one)”机制将多类学习问题分解为 $(q/2)$ 个二类学习问题，每个二类学习问题将一组标记配对 $(y_j, y_k)(j < k)$ 中的标记分别作为正类和负类。

在偏标记学习问题中，由于训练样本的真实标记未知，上述的“一对多”分解机制以及“一对一”分解机制均无法直接使用。基于上述考虑，通过对纠错输出编码(Error-correcting output codes, ECOC)方法进行扩展，提出了一种新的偏标记学习算法 PL-ECOC[29]。该算法不仅可以对偏标记训练样本进行学习，同时继承了二类分解机制在构建多类分类器时的简明特性。

在传统监督学习框架下，ECOC 通过特定的编码以及解码过程实现二类分解[30,31]。在编码阶段，ECOC 利用大小为 $q \times L$ 的二值编码矩阵 $\boldsymbol{M} \in \{-1, +1\}^{q \times L}$ 构建二类分类器。其中，矩阵的每一行 $\boldsymbol{M}(j,:)$ 代表与类别标记 y_j 对应的长度为 L 比特的码字；矩阵的每一列 $\boldsymbol{M}(:,l)$ 将标记空间 y 划分为互不相交的两个子集 y_l^+ 以及 y_l^-

$$y_l^+ = \{y_j \mid \boldsymbol{M}(j,l) = +1, 1 \leqslant j \leqslant q\} \tag{19}$$

$$y_l^- = \{y_j \mid \boldsymbol{M}(j,l) = -1, 1 \leqslant j \leqslant q\} \tag{20}$$

将隶属于 y_l^+ 与 y_l^- 的训练样本分别作为正样本与负样本提交给二类学习算法,即可获得二类分类器 $h_l: x \to \{-1,+1\}$。

在解码阶段,给定测试样本 x^*,基于各二类分类器在 x^* 上的分类输出可得与 x^* 对应的长度为 L 比特的码字:$\boldsymbol{h}(x^*)=[h_1(x^*),h_2(x^*),\cdots,h_L(x^*)]^{\mathrm{T}}$。ECOC 将码字与 $\boldsymbol{h}(x^*)$ 最接近的类别作为测试样本的预测输出

$$f(x^*)=\arg\max_{y_j\in y}\mathrm{dist}(\boldsymbol{h}(x^*),\boldsymbol{M}(j,:)) \tag{21}$$

距离函数 dist 可采用多种方式实现,例如:海明距离[30],欧氏距离[32],基于损失的距离[31,33]等。

为了扩展 ECOC 以求解偏标记学习问题,其关键在于如何使用偏标记训练集构建与编码矩阵各列对应的二类分类器。给定偏标记训练样本$(\boldsymbol{x}_i,S_i)$,PL-ECOC 算法不再试图对候选标记集合 S_i 进行消歧操作,而是将其作为一个整体进行处理。在构建各二类分类器 h_l 时,若候选标记集合 S_i 完全落在 y_l^+ 中,则将 $\boldsymbol{x}_i$ 作为正样本;若候选标记集合 S_i 完全落在 y_l^- 中,则将 $\boldsymbol{x}_i$ 作为负样本;否则,$\boldsymbol{x}_i$ 在构建 h_l 的过程中将被忽略。

算法 5 给出了基于非消歧策略的 PL-ECOC 算法[29]伪码描述。给定偏标记训练集 D,算法首先初始化编码矩阵(步骤 1);然后,算法基于编码矩阵的每一列元素将偏标记训练集转化为二类训练集(步骤 3~8)从而构建相应的二类分类器(步骤 9);最后,算法基于 ECOC 解码规则预测测试样本的概念标记(步骤 11~12)。

算法 5 PL-ECOC 算法[29]的伪码描述

$y^*=$PL-ECOC(D,L,B,x^*)

1. Randomly generate a $q\times L$ binary coding matrix $\boldsymbol{M}\in\{-1,+1\}^{q\times L}$;
2. FOR $l=1$ to L DO
3. Dichotomize the label space y into y_l^+ and y_l^- according to Eqs. (19) and (20);
4. Set $D_l=\varnothing$;
5. FOR $i=1$ to m DO
6. IF $S_i\subseteq y_l^+$ THEN $D_l=D\cup\{(\boldsymbol{x}_i,+1)\}$;
7. IF $S_i\subseteq y_l^-$ THEN $D_l=D\cup\{(\boldsymbol{x}_i,-1)\}$;
8. ENDFOR
9. Induce binary classifier $h_l\leftarrow B(D_l)$;
10. ENDFOR
11. Create $\boldsymbol{h}(x^*)$ by querying the binary classifiers $\boldsymbol{h}(x^*)=[h_1(x^*),h_2(x^*),\cdots,h_L(x^*)]^{\mathrm{T}}$;
12. Return $y^*=f(x^*)$ according to Eq. (21)。

3 结束语

偏标记学习是一类重要的弱监督机器学习框架,在该框架下,每个对象可能同时具有多个候选标记,但其中仅有一个为其真实标记。为了有效地对偏标记对象进行学习建模,本文分别对基于辨识消歧策略、基于平均消歧策略以及基于非消歧策略的偏标记学习算法进行了介绍。对于偏标记学习领域,认为有如下问题值得进一步深入研究:首先,一般而言,损失函数的设计体现了算法关于学习问题本质的刻画,是影响系统泛化性能的关键因素之一。在偏标记学习问题中,学习系统的目标是得到多类分类器 $f: x\to y$,对于分类系统其最常用的经验损失函数即为"0-1 损失":$\frac{1}{m}\sum_{i=1}^{m}[f(\boldsymbol{x}_i)\neq y_i]$。然而,由于各偏标记训练样本的真实标记 y_i 未知,必须对传统的 0-1 损失函数进行改造以适应偏标记学习问题的需要[34]。其次,当训练集中包含的监督信息有限时,充分挖掘其他潜在的有用信息显得十分重要,而相关

性信息是其中一种典型代表。一方面,样本之间通常满足流形假设,输入相近的样本其输出(候选标记集合)亦相近;另一方面,标记之间尤其是候选标记之间通常具有很强的关联,如语义相近的标记在候选标记集合中同时出现的可能性较高等。此外,为了提升偏标记学习系统的泛化性能,一种很自然的想法是融合相关弱监督学习框架的可用技术。例如,对于偏标记训练样本($\boldsymbol{x}_i$, S_i)而言,若 S_i 中含有的候选标记数较高($|S_i| \approx q$),则该样本在很大程度上可视为未标记样本(即任一标记均可能成为其真实标记),从而为半监督学习技术的引入提供了可能。考察偏标记学习框架与其他弱监督学习框架的适当融合也是一个值得尝试的研究方向。

参考文献:

[1] Mitchell T M. Machine learning[M]. New York: McGraw-Hill, 1997.

[2] Pfahringer B. Learning with weak supervision: Charting the territory[C]// Keynote Talk at the 1st International Workshop on Learning with Weak Supervision (LAWS'12, in conjunction with ACML'12). Singapore:[s. n.], 2012.

[3] Cour T, Sapp B, Taskar B. Learning from partial labels[J]. Journal of Machine Learning Research, 2011, 12: 1501-1536.

[4] Cour T, Sapp B, Jordan C, et al. Learning from ambiguous labeled images[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, FL:[s. n.], 2009: 919-926.

[5] Zeng Z, Xiao S, Jia K, et al. Learning by associating ambiguously labeled images[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, OR:[s. n.], 2013: 708-715.

[6] Jie L, Orabona F. Learning from candidate labeling sets[C]// Advances in Neural Information Processing Systems 23. Cambridge, MA: MIT Press, 2010: 1504-1512.

[7] Liu L, Dietterich T. A conditional multinomial mixture model for superset label learning[C]// Advances in Neural Information Processing Systems 25. Cambridge, MA: MIT Press, 2012: 557-565.

[8] Grandvalet Y. Logistic regression for partial labels[C]// Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Annecy, France:[s. n.], 2002: 1935-1941.

[9] Jin R, Ghahramani Z. Learning with multiple labels[C]// Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003: 897-904.

[10] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning[M]. Cambridge, MA: MIT Press, 2006.

[11] Zhu X, Goldberg A B. Introduction to semi-supervised learning[C]// Synthesis Lectures to Artificial Intelligence and Machine Learning. San Francisco, CA: Morgan & Claypool Publishers, 2009: 1-130.

[12] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data[C]// Data Mining and Knowledge Discovery Handbook. Berlin: Springer, 2010: 667-686.

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

[14] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1995, 2(1): 263-286.

[15] Amores J. Multiple instance classification: Review, taxonomy and comparative study[J]. Artificial Intelligence, 2013, 201: 81-105.

[16] Satoh S, Nakamura Y, Kanade T. Name-it: Naming and detecting faces in news videos[J]. IEEE Multimedia, 1999, 6(1): 22-35.

[17] Barnard K, Duygulu P, Forsyth D A, et al. Matching words and pictures[J]. Journal of Machine Learning Research, 2003, 3: 1107-1135.

[18] Berg T L, Berg A C, Edwards J, et al. Names and faces in the news[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC:[s. n.], 2004: 848-854.

[19] Everingham M, Sivic J, Zisserman A. "Hello! My name is… Buffy"—Automatic naming of characters in TV video[C]// Proceedings of the 17th British Machine Vision Conference. Edinburgh, UK:[s. n.], 2006: 889-908.

[20] Ramanan D, Baker S, Kakade S. Leveraging archival video for building face datasets[C]// Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil:[s. n.], 2007: 1-8.

[21] Nguyen N, Caruana R. Classification with partial labels[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV:[s. n.], 2008: 381-389.

[22] Hüllermeier E, Beringer J. Learning from ambiguously labeled examples[J]. Intelligent Data Analysis, 2006, 10(5): 419-439.
[23] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 1977, 39(1): 1-38.
[24] Della Pietra S, Della Pietra V, Lafferty J. Inducing features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4): 380-393.
[25] Koller D, Friedman N. Probabilistic graphical models: Principles and techniques[M]. Cambridge, MA: MIT Press, 2009.
[26] Crammer K. Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines[J]. Journal of Machine Learning Research, 2001, 2: 265-292.
[27] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM[C]// Proceedings of the 24th International Conference on Machine Learning. Corvallis, OR:[s. n.], 2007: 807-814.
[28] Aha D W. Special AI review issue on lazy learning[J]. Artificial Intelligence Review, 1997, 11(1-5):7-10.
[29] Zhang M L. Disambiguation-free partial label learning[C]// Proceedings of the 14th SIAM International Conference on Data Mining. Philadelphia, PA:[s. n.], 2014, 37-45.
[30] Dietterich T G, Bakiri G. Solving multiclass learning problem via error-correcting output codes[J]. Journal of Artificial Intelligence Research, 1995, 2(1): 263-286.
[31] Zhou Z H. Ensemble methods: Foundations and algorithms[M]. Boca Raton, FL: Chapman & Hall/CRC, 2012.
[32] Pujol O, Escalera S, Radeva P. An incremental node embedding technique for error correcting output codes[J]. Pattern Recognition, 2008, 41(2): 713-725.
[33] Allwein E L, Schapire R E, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers[J]. Journal of Machine Learning Research, 2000, 1: 113-141.
[34] Cid-Sueiro J. Proper losses for learning from partial labels[C]// Advances in Neural Information Processing Systems 25. Cambridge, MA: MIT Press, 2012: 1574-1582.

作者简介:**张敏灵**(1979-),男,教授,研究方向:机器学习、数据挖掘,E-mail:zhangml@seu.edu.cn。