软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2851−2864 [doi: 10.13328/j.cnki.jos.005339]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于标记与特征依赖最大化的弱标记集成分类*

谭桥宇[1], 余国先[1], 王 峻[1], 郭茂祖[2]

[1](西南大学 计算机与信息科学学院,重庆 400715)

[2](北京建筑大学 电气与信息工程学院,北京 100044)

通讯作者: 王峻, E-mail: kingjun@swu.edu.cn

摘 要: 弱标记学习是多标记学习的一个重要分支,近几年已被广泛研究并被应用于多标记样本的缺失标记补全和预测等问题.然而,针对特征集合较大、更容易拥有多个语义标记和出现标记缺失的高维数据问题,现有弱标记学习方法普遍易受这类数据包含的噪声和冗余特征的干扰.为了对高维多标记数据进行准确的分类,提出了一种基于标记与特征依赖最大化的弱标记集成分类方法 EnWL.EnWL 首先在高维数据的特征空间多次利用近邻传播聚类方法,每次选择聚类中心构成具有代表性的特征子集,降低噪声和冗余特征的干扰;再在每个特征子集上训练一个基于标记与特征依赖最大化的半监督多标记分类器;最后,通过投票集成这些分类器实现多标记分类.在多种高维数据集上的实验结果表明,EnWL 在多种评价度量上的预测性能均优于已有相关方法.

关键词: 弱标记学习;高维数据;特征子集;依赖最大化;集成分类

中图法分类号: TP181

中文引用格式: 谭桥宇,余国先,王峻,郭茂祖.基于标记与特征依赖最大化的弱标记集成分类.软件学报,2017,28(11):2851−2864. http://www.jos.org.cn/1000-9825/5339.htm

英文引用格式: Tan QY, Yu GX, Wang J, Guo MZ. Ensemble weak-label classification by maximizing dependency between label and feature. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2851−2864 (in Chinese). http://www.jos.org.cn/1000-9825/5339.htm

Ensemble Weak-Label Classification by Maximizing Dependency Between Label and Feature

TAN Qiao-Yu[1], YU Guo-Xian[1], WANG Jun[1], GUO Mao-Zu[2]

[1](College of Computer and Information Science, Southwest University, Chongqing 400715, China)

[2](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: Weak label learning is an important sub-branch of multi-label learning which has been widely studied and applied in replenishing missing labels of partially labeled instances or classifying new instances. However, existing weak label learning methods are generally vulnerable to noisy and redundant features in high-dimensional data where multiple labels and missing labels are more likely present. To accurately classify high-dimensional multi-label instances, in this paper, an ensemble weak label classification method is proposed by maximizing dependency between labels and features (EnWL for short). EnWL first repeatedly utilizes affinity propagation clustering in the feature space of high-dimensional data to find cluster centers. Next, it uses the obtained cluster centers to construct representative feature subsets and to reduce the impact of noisy and redundant features. Then, EnWL trains a semi-supervised multi-label classifier by maximizing the dependency between labels and features on each feature subset. Finally, it combines these base classifiers

* 基金项目: 国家自然科学基金(61402378, 61571163, 61532014, 61671189); 重庆市基础与前沿研究项目(cstc2014jcyjA40031, cstc2016jcyjA0351)

Foundation item: National Natural Science Foundation of China (61402378, 61571163, 61532014, 61671189); Natural Science Foundation of Chongqing Science and Technology Commission (cstc2014jcyjA40031, cstc2016jcyjA0351)

本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授推荐.

收稿时间: 2017-03-29; 修改时间: 2017-06-16; 采用时间: 2017-08-23

into an ensemble classifier via majority vote. Experimental results on several high-dimensional datasets show that EnWL significantly outperforms other related methods across various evaluation metrics.

Key words: weak label learning; high-dimensional data; feature subset; dependency maximization; ensemble classification

多标记学习问题是机器学习领域中一类重要的学习问题[1].在该学习问题中,每个训练样本可同时标注多个不同的类标记.例如,一个蛋白质可以同时拥有多个功能标记,如“新陈代谢”“信号传导”和“光合作用”等;一篇网页新闻可以同时拥有多个类别属性,如“政治”“经济”“文化”和“体育”.近年来,多标记学习方法已被成功应用于文本分类[2,3]、图像标注[4]和蛋白质功能预测[5,6]等任务中.早期对多标记学习问题的研究大多是在训练样本的标记完整的假设下进行[7,8],即假设每个样本的已有标记信息是完整的.但在很多现实世界的应用中,为每个多标记样本提供其对应的完整标记信息往往相当困难,而收集标记不完整的弱标记(weakly labeled)样本和大量未标记样本则相对较为容易.现有多标记学习方法由于忽视了已标注样本的标记缺失性,通常并不能基于弱标记样本进行有效地学习.弱标记学习(multi-label weak label learning)[9−11]考虑了多标记样本的标记信息存在缺失这一特点,它可以增量标注样本的缺失标记,也可以基于弱标记的多标记样本对新样本进行更准确的分类,已成为多标记学习的一个重要分支和研究热点[12−15].

近年来,随着移动互联网和高通量技术的不断发展应用,获取的数据越来越多,这些数据不仅规模日益增大,数据维度也日益增高,其包含的语义信息也更为丰富.对这类数据进行手工标注需要大量的人力物力,特别是当候选标记类别较多时,为每个样本提供完整的标记集合非常困难,甚至不可行.例如,完整标注一个高维图像样本集需要仔细查看每个图像中所有实体和实体间关联语义,再逐一标注确保无遗漏.因此,迫切需要研究针对复杂高维数据的弱标记学习方法.然而,现有弱标记学习方法并未对复杂高维数据进行特别处理.由于高维数据通常包含大量的冗余和噪声特征,这些特征破坏了高维数据之间距离度量的可靠性[16],已有的弱标记学习方法在高维数据上的预测精度因此降低.针对上述问题,本文提出一种基于标记与特征依赖最大化的弱标记集成分类方法(ensemble weak-label classification by maximizing dependency between label and feature,简称 EnWL). EnWL 的主要流程如下.

(1) 在不同参数设置下,多次利用邻接传播聚类方法 APC(affinity propagation clustering)[17]在高维数据的特征空间进行聚类,基于每次 APC 发现的簇中心构成具有代表性的特征子集.

(2) 基于每个特征子集将高维数据投影对应的低维空间,再训练基于标记与特征依赖最大化的半监督多标记分类器.

(3) 对这些分类器进行投票集成.

在多种数据集上与相关对比方法的实验结果比较表明,EnWL 在多种评价度量上均显著优于这些方法,APC 选择的代表性特征子集和最大化标记与特征之间的依赖均可提升弱标记分类的精度.

本文第 1 节将介绍弱标记学习的相关工作.第 2 节给出本文提出的方法.第 3 节汇报实验结果和对比分析.第 4 节总结全文.

1 弱标记学习的相关工作

限于篇幅,本节主要介绍弱标记学习的相关工作,关于多标记学习方面的研究,可以参考综述文献[1].

弱标记学习方法可以对弱标记样本的缺失标记进行补全,还可以直接利用弱标记样本对新样本的多个标记同时进行预测[9,15].从训练中是否利用未标记样本的角度,弱标记学习大致可以分为两类.

- 监督式弱标记学习方法[10,12,18,19]假定大量有标记样本可以轻易获得,只利用包含标记信息但存在缺失的弱标记样本进行训练.比如,Sun 等人[10]提出一种弱标记学习方法,该方法假设不同标记的样本之间存在较大的间隔,标记之间的关系可以通过一组基于低秩逼近的关联描述,但该方法需通过耗时的二次规划求取低秩逼近,难以处理较大的数据和标记集合;Yu 等人[12]提出一种经验风险最小化的学习框架对大规模弱标记学习问题进行研究;Wu 等人[18]基于标记一致性和标记平滑性假设提出一种监督式

弱标记学习方法;Bucak 等人[19]提出一种基于排序损失和组稀疏损失的多标记学习方法 MLR- GL.监督式弱标记学习方法通常需要利用大量弱标记样本才能获得较好的分类性能,然而在很多现实应用中,很难获取足够的弱标记样本,而获取大量的未标记样本相对容易.监督式弱标记学习方法由于未能利用大量的未标记样本,其学习性能往往受到限制.

- 半监督弱标记学习方法[11,20−22]综合利用少量弱标记样本和大量的未标记样本进行学习训练.比如,Wu 等人[11]提出一种基于标记一致性和标记平滑性的半监督弱标记学习方法 MLML;Zhao 等人[20]提出一种半监督弱标记学习方法,该方法结合流形正则项对未标记样本加以利用;Wu 等人[21]利用标记间的高阶关系,提出一种基于混合图的弱标记学习方法;Yu 等人[22]通过最大化标记与特征间的依赖,提出一种基于依赖最大化的弱标记学习方法 ProDM,并应用到蛋白质功能预测中,ProDM 既能对多标记样本的缺失标记进行补全,又能预测完全未标记样本的标记;Wu 等人[23]显式地考虑标记间的不平衡问题[24],提出了基于类别不平衡的弱标记学习方法 MMIB.

上述弱标记学习方法主要针对一般的弱标记学习,并未对复杂的高维数据进行特别处理,易受高维数据中噪声和冗余特征的影响.Kong 等人[25]首先利用 MDDM 算法[26]对高维数据进行降维,再在降维后的数据上基于局部平滑性假设提出一种直推式多标记分类方法 Tram.MDDM 通过最大标记与特征之间的依赖进行维数约减,但该方法假定所有标记都共享一个特征集合,忽视了不同标记之间可能拥有不同的关键特征集合.Zhang 等人[27]假设每个标记都有自己的关键特征集合,提出了基于标记关键特征的多标记学习方法 Lift,但当标记空间较大时,该方法为每个标记寻找关键特征需要大量的计算开销.

上述针对高维数据的多标记学习方法通常假定训练样本具有完整的标记信息,忽视了高维数据普遍存在的标记缺失,且易受噪声和冗余特征影响,预测精度有限.

针对弱标记学习在高维数据上有更普遍的需求,而已有方法易受高维数据中噪声和冗余特征的干扰等问题,本文提出一种基于标记与特征依赖最大化的弱标记集成分类方法 EnWL.EnWL 首先利用 APC 在高维数据的特征空间进行多次聚类,并利用聚类中心构成具有代表性的特征子集,降低噪声和冗余特征的干扰;其次,考虑到标记与特征信息的依赖关系,在每个特征子集上单独训练一个基于依赖最大化的半监督弱标记分类器,提高这些分类器的性能;最后,集成整合这些分类器实现弱标记集成分类.下文将对 EnWL 具体工作原理和流程进行详细描述.

2 基于依赖最大化的弱标记集成方法

假设 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N]\in\mathbb{R}^{D\times N}$ 为 N 个训练样本,其中,$\boldsymbol{x}_i\in\mathbb{R}^D$ 为第 i 个训练样本.$\boldsymbol{Y}=[\boldsymbol{y}^1,\boldsymbol{y}^2,\ldots,\boldsymbol{y}^N]\in\mathbb{R}^{C\times N}$ 为这 N 个训练样本的标记指示矩阵,其中,$\boldsymbol{y}_i\in\mathbb{R}^C$ 为 $\boldsymbol{x}_i$ 的标记向量,C 表示标记类别数.如果 $\boldsymbol{x}_i$ 属于标记 $c(1\leq c\leq C)$,则 $\boldsymbol{y}_{ic}=1$;否则,$\boldsymbol{y}_{ic}=0$.为不失一般性,假设 $N=l+u$ 个样本中,前 l 个样本为标记信息仅部分已知的弱标记样本,后 u 个样本的标记信息完全未知.弱标记分类的目标是:通过利用少量弱标记样本和大量未标记样本,学习到一个决策函数 $f(\boldsymbol{x})\rightarrow\{1,0\}^C$ 对新样本 $\boldsymbol{x}$ 的多个标记进行预测.

2.1 代表性特征子集获取方法

相对于低维数据,高维数据由于特征集合大,不仅更容易出现标记缺失现象,而且其中包含的大量噪声和冗余特征还会降低弱标记分类的性能.如果能够有效地降低噪声和冗余特征的干扰,则能够提高弱标记分类的预测精度.近年来,特征选择作为一种有效的剔除噪声和冗余特征的高维数据处理方法,受到研究者的广泛关注[28].例如,Dash 等人[29]利用 k-means 聚类方法在特征维度聚类,将 k 个簇中心选为特征子集并结合到后续学习任务中,提升了学习效果.为了获得较好的聚类结果,k-means 方法通常需要初始的簇中心尽可能地接近真实的簇中心和设置合适的聚类个数 k,然而在实际应用中,这些要求均不容易满足.此外,随着聚类规模的增大,该方法往往难以获得较好的聚类结果[17].针对这一问题,Frey 等人[17]提出了一种基于图的近邻传播聚类方法 APC,该方法通过近邻样本间信息迭代传播实现聚类.相对于 k-means,APC 由于只有 1 个输入参数 γ 且无需初始化簇中

心,使用更为方便[30].此外,APC 在大规模的样本空间上也能获得较好的聚类结果[17,30].基于此,考虑到高维数据中特征集合较大且通常含有噪声和冗余特征,本文采用 APC 在高维数据的特征空间而不是样本空间进行聚类,再利用聚类中心构成具有代表性的特征子集,进而降低噪声和冗余特征的干扰.由于篇幅限制,更多关于 APC 的描述可参考文献[17].为避免单次 APC 获取的代表性特征子集描述高维数据全局特征的局限性,本文对 APC 的参数γ设置不同的值,获取到不同的聚类中心,进而构造了多个具有代表性的特征子集.本文基于代表性特征子集生成低维数据集的方法分为以下两步.

1) 在一定范围内选取 T 个不同$\gamma_t(1\leq t\leq T)$值,其中,γ_t 表示第 t 个 APC 的参数值.记 T 个特征子集对应的聚类中心个数分别为$\{d_1,d_2,\ldots,d_T\}(d_t<<D)$.假设 $\boldsymbol{r}_t$ 代表第 t 个特征子集的指示向量,$\boldsymbol{r}_t$ 满足:

$$\sum_{j=1}^{D}\boldsymbol{r}_{tj}=d_t \tag{1}$$

其中,$\boldsymbol{r}_{tj}=1$ 表示第 j 个特征在第 t 个特征子集中,$\boldsymbol{r}_{tj}=0$ 表示该特征不在这个特征子集中.

2) 产生第 t 个特征子集对应的低维投影数据集:

$$\boldsymbol{Z}_t=X(\boldsymbol{r}_t) \tag{2}$$

其中, $\boldsymbol{Z}_t=[\boldsymbol{z}_t^1,\boldsymbol{z}_t^2,...,\boldsymbol{z}_t^N]\in\mathbb{R}^{d_t\times N}.\boldsymbol{z}_t^i\in\mathbb{R}^{d_t}$ 代表 $\boldsymbol{Z}_t$ 中的第 i 个样本,它实际为 $\boldsymbol{x}_i$ 在第 t 个 APC 获取的特征子集上的投影.基于 T 个不用的γ值,便可获得 T 个不同的低维数据集 $Z=\{\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_T\}$.

2.2 基于标记与特征依赖最大化的弱标记集成分类方法

本节主要介绍本文提出的基于标记和特征依赖最大化的弱标记集成分类方法 EnWL.EnWL 的基础分类器最小化的目标方程形式如下.

$$\Psi(f)=\Omega_1(f,\boldsymbol{x},\boldsymbol{y})+\alpha\Omega_2(f)+\beta\Omega_3(f) \tag{3}$$

上式的第 1 项为在 l 个弱标记样本的经验损失,第 2 项为综合利用 l 个弱标记样本和μ个未标记样本的正则化,第 3 项度量预测的样本标记与样本特征间的依赖.参数α和β用于调节这 3 项的相对重要性.

令 $f(\boldsymbol{x})=\boldsymbol{P}^T\boldsymbol{x}(\boldsymbol{P}\in\mathbb{R}^{D\times C})$为样本 $\boldsymbol{x}$ 预测的标记向量,公式(3)中的第 1 项定义为

$$\Omega_1(f,\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{l}\|\boldsymbol{P}^T\boldsymbol{x}_i-\boldsymbol{y}_i\|^2 \tag{4}$$

其中,$\boldsymbol{y}_i$ 为弱标记样本 $\boldsymbol{x}_i$ 的标记向量.公式(4)最小化弱标记样本的经验损失是基于一致性假设[31],即弱标记样本的预测结果应该与其已知标记一致.

公式(4)忽视了前 l 个弱标记样本的标记存在缺失的特点.多标记样本的标记间存在关联关系,大量研究表明,在多标记学习中,合理地利用标记间的关联信息,可以提高多标记学习的性能[1].已有多种方法应用于衡量标记间关联性并成功结合到多标记学习中[22,31,32],余弦相似度量由于其简单、直观而被广泛应用[22],因此,本文基于余弦相似性度量定义标记间的关联矩阵 $\boldsymbol{M}\in\mathbb{R}^{C\times C}$ 如下:

$$\boldsymbol{M}(c_1,c_2)=\frac{\boldsymbol{Y}_{c_1}.^T\boldsymbol{Y}_{c_2}.}{\|\boldsymbol{Y}_{c_1}.\|\|\boldsymbol{Y}_{c_2}.\|} \tag{5}$$

其中,$\boldsymbol{M}(c_1,c_2)$表示标记 c_1 和 c_2 之间的相关性大小,$\boldsymbol{Y}_{c_1}.$ 为 $\boldsymbol{Y}$ 的第 c_1 行.从公式(5)可知,当两个标记同时标注的样本个数越多时,它们之间的关联($\boldsymbol{M}(c_1,c_2)$)就越大.借鉴 Kong 等人[9]和 Yu 等人[22]的工作,本文利用标记间的关联矩阵 $\boldsymbol{M}$ 对弱标记样本 $\boldsymbol{x}_i$ 的缺失标记进行初步预估,方式如下:

$$\tilde{\boldsymbol{y}}_{ic}=\begin{cases}\boldsymbol{y}_i^T\boldsymbol{M}(\cdot,c), & \text{if } \boldsymbol{y}_{ic}=0\\ 1, & \text{other}\end{cases} \tag{6}$$

$\tilde{\boldsymbol{y}}_i$ 表示对弱标记样本 $\boldsymbol{x}_i$ 的缺失标记补充后得到的标记向量.公式(6)通过弱标记样本已知的标记和标记间的关联关系对其缺失标记进行估计.如若已知 $\boldsymbol{y}_{ic}=0$,且标记 c 与该样本已标注的标记间有较大的相关性,则标记 c 很可能是缺失标记,$\tilde{\boldsymbol{y}}_{ic}$ 将被赋予一个较大的概率值.换句话说,如果一张图片已标注了“海鸥”,则“大海”和“岛屿”比“老虎”和“草原”更可能为该图片的缺失标记.为保证 $\tilde{\boldsymbol{y}}_{ic}\in[0,1]$,当 $\tilde{\boldsymbol{y}}_{ic}>0$ 时,将 $\tilde{\boldsymbol{y}}_{ic}$ 归一化为 $\tilde{\boldsymbol{y}}_{ic}/\|\tilde{\boldsymbol{y}}_i\|$.在考

虑前 l 个弱标记样本标记缺失的基础上,公式(4)可重新定义为

$$\Omega_1(f,\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{l}\|\boldsymbol{P}^T\boldsymbol{x}_i-\tilde{\boldsymbol{y}}_i\|^2=\sum_{i=1}^{N}((\boldsymbol{P}^T\boldsymbol{x}_i-\tilde{\boldsymbol{y}}_i)\boldsymbol{V}_{ii}(\boldsymbol{P}^T\boldsymbol{x}_i-\tilde{\boldsymbol{y}}_i)^T)=tr((\boldsymbol{P}^T\boldsymbol{X}-\tilde{\boldsymbol{Y}})\boldsymbol{V}(\boldsymbol{P}^T\boldsymbol{X}-\tilde{\boldsymbol{Y}})^T) \tag{7}$$

其中,$\tilde{\boldsymbol{Y}}=[\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2,...,\tilde{\boldsymbol{y}}_N]$;$tr()$表示矩阵的迹运算;$\boldsymbol{V}\in\mathbb{R}^{N\times N}$是对角矩阵,其前 l 个对角元素为 1,其余对角元素为 0.综合利用少量有标记样本和大量未标记样本,可以提高学习器的预测性能[33].为了综合利用少量弱标记样本和大量未标记样本,本文基于 N 个样本构造一个 k 近邻图 $\boldsymbol{W}$,图中的每个节点代表一个样本,边的权重代表样本之间的相似度,$\boldsymbol{W}$ 的计算方式如下:

$$\boldsymbol{W}_{ij}=\begin{cases}1, & \boldsymbol{x}_i\in kNN(\boldsymbol{x}_j)\text{ or }\boldsymbol{x}_j\in kNN(\boldsymbol{x}_i)\\ 0, & \text{others}\end{cases} \tag{8}$$

其中,$\boldsymbol{W}_{ij}$ 表示样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的相似度,$\boldsymbol{x}_i\in kNN(\boldsymbol{x}_j)$表示 $\boldsymbol{x}_i$ 属于 $\boldsymbol{x}_j$ 基于欧氏距离的 k 近邻之一.为了简便,公式(8)中仅考虑 0-1 权重.基于平滑性假设[34],即相似样本通常拥有相似的标记,公式(3)的第 2 项$\Omega_2(f)$可定义为

$$\begin{aligned}\Omega_2(f)&=\frac{1}{2}\sum_{i,j=1}^{N}\|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|^2\boldsymbol{W}_{ij}=\frac{1}{2}\sum_{i,j=1}^{N}\|\boldsymbol{P}^T\boldsymbol{x}_i-\boldsymbol{P}^T\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}\\&=tr\left(\boldsymbol{P}^T\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{W}_{ii}\boldsymbol{x}_i^T)\boldsymbol{P}-\boldsymbol{P}T\left(\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{W}_{ij}\boldsymbol{x}_j^T)\boldsymbol{P}\right)\right)\\&=tr(\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{\Lambda}-\boldsymbol{W})\boldsymbol{X}^T\boldsymbol{P})=tr(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T\boldsymbol{P})\end{aligned} \tag{9}$$

其中,$\boldsymbol{\Lambda}$为对角矩阵,$\boldsymbol{\Lambda}_{ii}=\sum_{j=1}^{N}\boldsymbol{W}_{ij}$;$\boldsymbol{L}=\boldsymbol{\Lambda}-\boldsymbol{W}$为 $\boldsymbol{W}$ 对应近邻图上的图 Laplacian 矩阵[35].最小化公式(9),可以对弱标记样本的缺失标记作进一步补充.例如:若弱标记样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 互为近邻样本且 $\boldsymbol{W}_{ij}=1$,若已知 $\boldsymbol{y}_{ic_1}=1,\boldsymbol{y}_{jc_2}=1$,则标记 c_2 很可能为样本 $\boldsymbol{x}_i$ 的缺失标记,标记 c_1 也很可能为样本 $\boldsymbol{x}_j$ 的缺失标记.公式(9)不仅可以综合利用 N 个样本的分布信息,还可以对 u 个完全未标注样本的标记进行预测.

样本的标记依赖于样本的特征信息[22,26,36],例如,若某张图片被标注为“老虎”或“狮子”,则这张图片中一定有特定的区域(特征)与“老虎”或“狮子”相关.为利用样本标记对样本特征的依赖,本文采用 HSIC(Hilbert-Schmidt independence criterion)[37]对标记和特征间的依赖关系进行量化描述,并定义公式(3)的第 3 项为

$$\Omega_3(f)=(N-1)^{-2}tr(\boldsymbol{HKHS}) \tag{10}$$

其中,$\boldsymbol{H},\boldsymbol{K},\boldsymbol{E}\in\mathbb{R}^{N\times N}$.$\boldsymbol{K}_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 基于特征的相似度,$\boldsymbol{S}_{ij}=s(f(\boldsymbol{x}_i),f(\boldsymbol{x}_j))$表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 上预测的标记向量间的相似度.$\boldsymbol{H}_{ij}=\boldsymbol{\delta}_{ij}-1/N$.若 $i=j$,则$\boldsymbol{\delta}_{ij}=1$;否则,$\boldsymbol{\delta}_{ij}=0$.需要指出的是,EnWL 最大化标记与样本特征之间的依赖也利用了标记间的关联性.本文设置 $\boldsymbol{K}=\boldsymbol{W},\boldsymbol{E}=\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{X}$.最大化公式(10)的目的是:当 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的特征信息很相似(即 $\boldsymbol{K}_{ij}$ 较大)时,则在这 2 个样本上的预测标记向量也应该很相似(即 $\boldsymbol{S}_{ij}$ 较大).实际上,$\boldsymbol{S}$ 可以看作样本之间的语义相似度,样本间语义相似度与样本之间的特征相似度通常正相关[38],语义相似度已被成功应用于多标记图像标注[39]和蛋白质功能预测[40]等领域.最大化多标记样本的标记信息与特征信息间的依赖可以在公式(7)缺失标记补充和公式(9)中平滑性假设的基础上,从样本间语义的角度进一步提高弱标记学习的性能.

在上述分析的基础上,公式(3)可重写为

$$\Psi(f)=tr((\boldsymbol{P}^T\boldsymbol{X}-\tilde{\boldsymbol{Y}})\boldsymbol{V}(\boldsymbol{P}^T\boldsymbol{X}-\tilde{\boldsymbol{Y}})^T)+\alpha tr(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T\boldsymbol{P})-\beta tr(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{P}) \tag{11}$$

对公式(11)求关于 $\boldsymbol{P}$ 的导数:

$$\frac{\partial\Psi(f)}{\partial\boldsymbol{P}}=2\boldsymbol{X}\boldsymbol{V}(\boldsymbol{X}^T\boldsymbol{P}-\tilde{\boldsymbol{Y}}^T)+2\alpha\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T\boldsymbol{P}-2\beta\boldsymbol{X}\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{P} \tag{12}$$

令 $\frac{\partial\Psi(f)}{\partial\boldsymbol{P}}=0$,可得:

$$\boldsymbol{P}=(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}^T+\alpha\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T-\beta\boldsymbol{X}\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{V}\tilde{\boldsymbol{Y}}^T \tag{13}$$

从公式(13)可以看出,当 $\boldsymbol{Y}$ 已知时,$\boldsymbol{P}$ 由 $\boldsymbol{L}$ 和 $\boldsymbol{W}$ 共同决定,而 $\boldsymbol{L}$ 也是基于 $\boldsymbol{W}$ 计算获得,因此公式(13)的预测投影向量 $\boldsymbol{P}$ 依赖于 $\boldsymbol{W}$.然而,由于高维数据中大量噪声和冗余特征的干扰,直接在原始特征空间定义的 $\boldsymbol{W}$ 通常并不能比较准确地描述样本之间的相似性[16],为此,本文提出的 EnWL 在上一节中每次 APC 获取的特征子集投影的

低维数据 $\boldsymbol{Z}_t$ 上计算 $\boldsymbol{W}$,求解对应 $\boldsymbol{P}_t$($\boldsymbol{P}_t \in \mathbb{R}^{d_t \times C}$),再通过下式进行投票集成:

$$f_{ens}(\boldsymbol{x}_i) = \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{x}_i^t) \tag{14}$$

其中, $f(\boldsymbol{x}_i^t) = \boldsymbol{P}_t^T \boldsymbol{x}_i^t$ 为预测的 $\boldsymbol{z}_t^i$ 属于 C 个不同标记的概率值向量,$f_{ens}(\boldsymbol{x}_i)$为 T 个如公式(11)中的基础分类器在样本 $\boldsymbol{x}_i$ 上的集成预测结果.基础分类器的精度以及它们之间的差异性决定了集成学习的效果[41],基础分类器很难同时达到精度和差异性的最大化.EnWL 通过在多个特征子集投影的低维数据上训练基于标记与特征依赖最大化的弱标记分类器,保证基础分类器的精度,同时,通过不同参数设置下的 APC 获取不同的特征子集,进而获得具有一定差异性的基础分类器,从而,EnWL 能够获得比这些基础分类器更高的精度.此外,这种集成策略还可以避免单个 APC 获取的特征子集刻画高维数据特征信息的不足,和降低 APC 对γ参数输入的依赖,提升 EnWL 的鲁棒性.

3 实验结果

3.1 实验数据

本文在 6 个多标记数据集上开展实验,其中,Arts[42],Reference[42]和 Eurlex[43]来自文本分类,ESPGame[44],Core15k[45]和 IAPRTC-12[46]来自图像分类.与文献[19]和文献[38]类似,本文剔除掉 ESPGame 中标记个数少于 5 的样本.表 1 给出这些数据集的统计信息,可以看出,这 6 个数据集的维度均大于 1 000.其中,ESPGame,Eurlex,Core15k 和 IAPRTC-12 的样本数目和样本维度之比接近 3:1,而 Arts 和 Reference 样本数目和样本维度之比为 1:3,在这些数据集上的分类学习任务普遍面临着维数灾难问题.

Table 1 Experimental datasets (Avg. is the average number of labels of an instance)

表 **1** 实验数据集(Avg.对应每个样本相关标记的平均个数)

数据集	样本数目 N	样本维度 D	标记个数 C	Avg.
ESPGame	10 457	3 000	268	6.41
Eurlex	19 348	5 000	201	2.21
Arts	7 441	17 973	19	1.62
Reference	7 929	26 397	15	1.15
Core15k	4 395	1 000	260	3.61
IAPRTC-12	12 985	5 184	291	7.07

3.2 对比算法及评价准则

为了分析比较 EnWL 算法的性能,本文将 MLR-GL[19],MLML[10],ProDM[22],MMIB[23],Tram[25]和 Lift[27]作为对比算法.其中,前 4 个是弱标记学习方法,后两个是基于特征选择的多标记学习方法.这些对比算法已在相关工作中做了详细介绍.MLML,ProDM,MMIB,Tram 是直推式(transductive)分类方法,只能预测训练集中未标记样本的标记,不能对训练集以外的新样本进行预测.为了将这些方法扩展为传导式(inductive)方法进行对比实验,本文通过设置新样本的标记为训练集中与其距离最近样本的标记,而训练集中未标记样本的标记基于它们各自的直推分类获得[8].实验中,本文依据原文建议的参数值范围和方式选取对比算法的最优参数进行实验,并在训练集上利用 5 重交叉验证优化 EnWL 的参数 T,α和β.T 的范围是 1~25,步长为 1;α和β的范围是 0.01~1,步长为 0.01;最终设置 T=15,α=0.9 和β=0.01.APC 对应不同的γ值通常产生不同的聚类个数,本文实验设置 APC 聚类个数为[100,500],以此范围搜索对应的γ值区间,在此区间随机产生 T 个不同的γ值,进而获取 T 个代表性特征子集.

弱标记学习的性能可以从多个不同的角度进行评价,为了综合评价性能,本文选用 HammLoss,RankLoss,AvgPrec,Coverage 和 AUC 这 5 种常用的多标记学习评价度量来衡量上述对比算法和 EnWL 的性能.其中,前 4 种度量的定义可参见文献[1].AUC 首先将每个测试样本的预测向量中的元素从大至小排序,在控制每个样本预测的标记个数从 1 增至 C 的同时,计算这些样本上对应的真阳性率和假阳性率;最后计算真阳性率-假阳性率对应曲线下的面积,以此面积大小衡量分类性能.AUC 的具体定义可参见文献[19].RankLoss,HammLoss 和 Coverage 的值越小,表示预测精度越高.为保持一致性,实验中汇报的是 1–RankLoss 和 1–HammLoss 的

值.Coverage 的值通常大于 1,故不作类似处理.这些度量从不同的方面评测多标记学习的性能,一种算法很难在所有度量上超过另一种算法.

3.3 实验结果

对每个数据集,本文分别随机选取 30%样本作为有标记的训练样本和测试样本,剩下 40%作为未标记的训练样本.本文假设这些数据集中样本的标记信息是完整的,在每次实验中,对每个样本的已有标记进行随机隐藏构造弱标记样本.为了方便表示,实验中用 m 表示一个弱标记样本的缺失标记数量,例如,m=3 表明该样本有 3 个标记被隐藏(缺失).如果一个样本的标记个数不足 m,本文不会将其所有标记隐藏,而是保证该样本至少有 1 个标记.为了减少随机性的影响,在每个数据集上每个算法重复 10 次独立随机实验,并记录每个对比算法在给定 m 下的 10 次平均结果.表 2~表 7 分别给出了这些算法在 ESPGame,Reference,Eurlex,Arts,Core15k 和 IAPRTC-12 上的实验结果(均值±方差),其中,↓表示该度量值越小,对应的结果越好.表中加粗的结果表明其在配对检验(95%置信度)中显著优于其他结果,或与最优结果之间无显著性差异.

从这些表中的结果可以观察到,EnWL 在绝大多数情况下都能获得比其他对比算法更好的结果.在 6 个数据集上的 90 次对比实验中,EnWL 的结果一直优于 MLR-GL 和 ProDM.EnWL 在 90 次对比实验中分别与 MLML,MMIB 和 Tram 获得了 7,3,13 次相似的最优结果,但在其他对比实验中,EnWL 均超过后者.

Table 2 Experimental results of EnWL and compared methods on ESPGame

表 **2** ESPGame 上本文 EnWL 方法与对比方法的实验结果

Metric	m	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.970±0.000**	0.948±0.000	0.957±0.001	0.952±0.000	0.951±0.000	0.951±0.000	0.958±0.000
	2	**0.969±0.000**	0.949±0.000	0.956±0.001	0.952±0.000	0.948±0.000	0.951±0.000	0.959±0.000
	3	**0.969±0.000**	0.949±0.001	0.953±0.001	0.951±0.000	0.940±0.000	0.951±0.000	0.958±0.000
1−RankLoss	1	**0.772±0.001**	0.684±0.016	0.769±0.002	0.643±0.024	0.768±0.003	0.675±0.049	0.768±0.000
	2	**0.771±0.001**	0.679±0.015	0.765±0.003	0.640±0.018	0.767±0.002	0.701±0.017	0.768±0.001
	3	**0.770±0.001**	0.671±0.020	0.757±0.003	0.638±0.005	0.770±0.001	0.661±0.059	0.767±0.000
AvgPrec	1	0.188±0.000	0.104±0.001	0.183±0.012	0.131±0.008	0.112±0.006	0.117±0.008	**0.207±0.000**
	2	0.169±0.010	0.104±0.003	0.163±0.015	0.132±0.006	0.121±0.005	0.116±0.005	**0.210±0.000**
	3	0.147±0.010	0.105±0.008	0.132±0.011	0.104±0.004	0.127±0.004	0.120±0.009	**0.208±0.001**
AUC	1	**0.777±0.001**	0.705±0.016	0.774±0.002	0.725±0.014	0.773±0.003	0.729±0.019	0.773±0.000
	2	**0.776±0.001**	0.700±0.015	0.769±0.003	0.723±0.009	0.772±0.002	0.738±0.006	0.772±0.001
	3	**0.773±0.001**	0.690±0.019	0.762±0.003	0.718±0.002	0.772±0.001	0.722±0.024	0.771±0.000
Coverage↓	1	**138.437±0.466**	218.707±2.983	151.389±0.618	192.442±1.651	149.152±0.304	158.493±4.008	156.111±0.312
	2	**138.491±0.205**	214.558±2.576	152.281±0.790	193.025±1.663	150.904±0.497	157.947±1.662	157.241±0.572
	3	**139.296±0.474**	217.138±3.207	154.696±1.094	202.554±1.336	150.037±0.436	162.974±4.376	156.196±0.161

Table 3 Experimental results of EnWL and compared methods on Reference

表 **3** Reference 上本文 EnWL 方法与对比方法的实验结果

Metric	m	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.895±0.015**	0.829±0.004	0.862±0.001	0.875±0.006	0.869±0.000	0.879±0.004	0.889±0.001
	2	**0.887±0.002**	0.818±0.005	0.863±0.001	0.876±0.000	0.870±0.001	0.881±0.006	0.884±0.001
	3	**0.885±0.001**	0.825±0.005	0.863±0.000	0.871±0.005	0.869±0.000	0.868±0.005	0.856±0.001
1−RankLoss	1	**0.784±0.008**	0.753±0.008	0.775±0.004	0.732±0.045	0.779±0.001	0.762±0.025	0.768±0.004
	2	**0.786±0.001**	0.748±0.037	0.768±0.004	0.728±0.003	**0.785±0.002**	0.760±0.042	0.762±0.002
	3	**0.784±0.002**	0.736±0.046	0.777±0.002	0.721±0.032	0.780±0.002	0.750±0.031	0.761±0.003
AvgPrec	1	**0.641±0.066**	0.506±0.043	0.562±0.004	0.328±0.014	0.574±0.003	0.378±0.010	0.407±0.005
	2	**0.640±0.006**	0.366±0.063	0.561±0.005	0.350±0.011	0.581±0.004	0.349±0.010	0.392±0.004
	3	**0.639±0.005**	0.426±0.067	0.563±0.001	0.328±0.012	0.574±0.001	0.343±0.004	0.348±0.003
AUC	1	**0.823±0.027**	0.769±0.005	0.780±0.003	0.764±0.012	0.792±0.001	0.811±0.004	0.818±0.004
	2	**0.815±0.041**	0.773±0.034	0.779±0.003	0.785±0.002	0.797±0.003	0.804±0.010	0.809±0.003
	3	**0.814±0.029**	0.726±0.043	0.781±0.002	0.778±0.010	0.792±0.002	0.782±0.006	0.791±0.003
Coverage↓	1	**3.019±0.141**	4.415±0.115	3.737±0.048	6.172±0.587	3.496±0.008	4.373±0.171	3.810±0.070
	2	**3.512±0.174**	5.094±0.576	3.854±0.060	5.178±0.061	**3.432±0.145**	4.966±0.467	4.348±0.036
	3	**3.567±0.162**	5.190±0.736	3.700±0.038	5.740±0.423	**3.504±0.142**	5.376±0.301	4.348±0.061

Table 4 Experimental results of EnWL and compared methods on Eurlex

表 4 Eurlex 上本文 EnWL 方法与对比方法的实验结果

Metric	*m*	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.978±0.002**	0.963±0.000	0.969±0.000	**0.978±0.000**	0.899±0.001	**0.979±0.002**	0.971±0.001
	2	**0.977±0.000**	0.963±0.000	0.968±0.001	**0.979±0.002**	0.890±0.003	**0.977±0.001**	0.973±0.001
	3	**0.974±0.001**	0.963±0.001	0.966±0.002	**0.975±0.001**	0.888±0.002	**0.975±0.001**	0.970±0.003
1−RankLoss	1	**0.934±0.001**	0.778±0.002	0.843±0.002	0.922±0.001	0.786±0.006	0.877±0.053	0.875±0.001
	2	**0.934±0.001**	0.789±0.002	0.824±0.001	0.923±0.000	0.773±0.006	0.876±0.001	0.869±0.006
	3	**0.931±0.001**	0.792±0.002	0.804±0.007	0.919±0.001	0.765±0.002	0.873±0.000	0.867±0.002
AvgPrec	1	**0.594±0.056**	0.144±0.006	0.322±0.002	0.569±0.006	0.209±0.002	0.574±0.012	0.457±0.002
	2	**0.591±0.002**	0.148±0.002	0.271±0.002	0.583±0.000	0.145±0.004	0.569±0.009	0.455±0.036
	3	**0.585±0.000**	0.152±0.007	0.242±0.003	0.569±0.003	0.139±0.012	0.564±0.003	0.450±0.079
AUC	1	**0.931±0.003**	0.792±0.001	0.862±0.002	**0.931±0.001**	0.811±0.003	0.927±0.019	0.876±0.000
	2	**0.924±0.007**	0.797±0.001	0.835±0.002	**0.925±0.000**	0.799±0.005	0.917±0.001	0.862±0.005
	3	0.923±0.006	0.797±0.003	0.813±0.006	**0.928±0.001**	0.792±0.003	0.904±0.000	0.862±0.001
Coverage↓	1	15.774±0.461	40.755±0.285	30.117±0.318	**14.913±0.126**	39.183±0.734	20.924±0.325	26.784±0.206
	2	17.619±0.882	39.350±0.206	34.486±0.293	**15.466±0.141**	41.017±0.899	21.918±0.130	28.807±0.893
	3	18.124±0.756	39.341±0.200	38.622±1.095	**16.310±0.024**	42.289±0.090	22.114±0.066	29.237±0.001

Table 5 Experimental results of EnWL and compared methods on Arts

表 5 Arts 上本文 EnWL 方法与对比方法的实验结果

Metric	*m*	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.886±0.007**	0.859±0.001	0.853±0.007	0.851±0.001	0.862±0.002	0.869±0.009	0.872±0.000
	2	**0.886±0.012**	0.859±0.000	0.846±0.006	0.849±0.003	0.861±0.002	0.870±0.006	0.870±0.002
	3	**0.884±0.016**	0.859±0.000	0.847±0.003	0.848±0.002	0.859±0.004	0.868±0.004	0.870±0.001
1−RankLoss	1	**0.822±0.032**	0.770±0.002	0.738±0.022	0.616±0.010	0.733±0.033	0.736±0.021	0.783±0.003
	2	**0.818±0.018**	0.770±0.001	0.717±0.036	0.593±0.014	0.693±0.022	0.716±0.021	0.783±0.002
	3	**0.811±0.006**	0.768±0.002	0.700±0.006	0.600±0.018	0.679±0.025	0.707±0.010	0.778±0.001
AvgPrec	1	**0.600±0.044**	0.437±0.006	0.396±0.039	0.374±0.009	0.320±0.026	0.384±0.009	0.539±0.003
	2	**0.601±0.061**	0.440±0.005	0.361±0.042	0.349±0.003	0.282±0.019	0.410±0.017	0.541±0.008
	3	**0.580±0.081**	0.431±0.006	0.352±0.023	0.387±0.017	0.293±0.030	0.350±0.009	0.525±0.006
AUC	1	**0.806±0.027**	0.765±0.002	0.733±0.020	0.751±0.006	0.779±0.026	0.735±0.013	0.761±0.002
	2	**0.795±0.021**	0.763±0.002	0.725±0.030	0.736±0.012	0.759±0.018	0.734±0.015	0.757±0.001
	3	**0.795±0.025**	0.761±0.001	0.703±0.005	0.730±0.012	0.748±0.022	0.727±0.010	0.755±0.003
Coverage↓	1	**4.683±0.430**	5.522±0.058	6.107±0.405	6.294±0.129	6.301±0.560	6.429±0.159	5.484±0.038
	2	**4.810±0.480**	5.535±0.024	6.438±0.631	6.594±0.142	7.192±0.311	6.443±0.285	5.434±0.020
	3	**4.903±0.292**	5.575±0.042	6.783±0.062	6.458±0.286	7.380±0.517	7.058±0.279	5.559±0.041

Table 6 Experimental results of EnWL and compared methods on Core15k

表 6 Core15k 上本文 EnWL 方法与对比方法的实验结果

Metric	*m*	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.953±0.000**	0.946±0.001	0.950±0.000	**0.953±0.000**	0.949±0.000	0.952±0.000	0.952±0.000
	2	**0.953±0.000**	0.949±0.000	0.950±0.000	**0.953±0.000**	0.941±0.000	0.951±0.000	0.951±0.000
	3	0.952±0.000	0.949±0.000	0.950±0.000	**0.953±0.000**	0.935±0.000	0.951±0.000	0.951±0.000
1−RankLoss	1	**0.835±0.002**	0.749±0.002	0.779±0.002	0.801±0.005	0.789±0.004	0.818±0.002	0.817±0.002
	2	**0.835±0.004**	0.745±0.001	0.742±0.002	0.813±0.003	0.777±0.006	0.815±0.001	0.801±0.002
	3	**0.828±0.000**	0.749±0.002	0.747±0.001	0.789±0.003	0.758±0.002	0.810±0.001	0.788±0.001
AvgPrec	1	**0.341±0.003**	0.216±0.006	0.273±0.001	0.223±0.001	0.202±0.003	0.248±0.001	0.305±0.001
	2	**0.333±0.008**	0.235±0.004	0.243±0.003	0.231±0.006	0.201±0.005	0.252±0.002	0.278±0.003
	3	**0.326±0.001**	0.234±0.006	0.239±0.002	0.229±0.003	0.191±0.004	0.256±0.002	0.278±0.002
AUC	1	**0.838±0.001**	0.753±0.002	0.786±0.001	0.831±0.003	0.797±0.005	0.836±0.002	0.821±0.001
	2	**0.837±0.005**	0.747±0.001	0.749±0.002	0.822±0.003	0.784±0.006	0.824±0.001	0.806±0.002
	3	**0.833±0.025**	0.750±0.002	0.747±0.002	0.815±0.002	0.762±0.003	0.815±0.001	0.791±0.002
Coverage↓	1	**38.945±0.186**	44.539±0.221	48.225±0.479	**38.694±0.309**	40.664±0.543	**38.022±0.493**	42.478±0.479
	2	**38.612±0.313**	46.132±0.497	55.727±0.242	**39.265±0.675**	43.608±0.591	41.446±0.304	45.901±0.242
	3	**39.199±0.337**	45.348±0.571	55.707±0.601	42.229±0.541	46.84±0.482	43.060±0.430	48.656±0.601

Table 7 Experimental results of EnWL and compared methods on IAPRTC-12

表 7 IAPRTC-12 上本文 EnWL 方法与对比方法的实验结果

Metric	*m*	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
1−HammLoss	1	**0.962±0.000**	0.949±0.000	0.959±0.000	0.953±0.000	0.943±0.000	0.954±0.000	0.950±0.006
	2	**0.961±0.000**	0.949±0.000	0.959±0.000	0.953±0.000	0.943±0.000	0.954±0.000	0.949±0.011
	3	**0.961±0.000**	0.949±0.000	0.959±0.000	0.953±0.000	0.942±0.000	0.954±0.000	0.949±0.010
1−RankLoss	1	**0.828±0.001**	0.718±0.021	0.801±0.000	0.775±0.003	0.750±0.002	0.801±0.001	0.740±0.008
	2	**0.822±0.001**	0.690±0.013	0.796±0.002	0.767±0.073	0.735±0.005	0.802±0.001	0.713±0.017
	3	**0.816±0.001**	0.679±0.016	0.798±0.001	0.688±0.001	0.727±0.004	0.800±0.000	0.682±0.014
AvgPrec	1	**0.280±0.001**	0.044±0.004	0.226±0.000	0.175±0.001	0.070±0.001	0.180±0.001	0.098±0.008
	2	**0.274±0.001**	0.042±0.002	0.224±0.001	0.175±0.004	0.068±0.003	0.180±0.002	0.090±0.018
	3	**0.270±0.001**	0.043±0.001	0.224±0.001	0.171±0.001	0.066±0.002	0.179±0.002	0.088±0.017
AUC	1	**0.829±0.001**	0.716±0.019	0.805±0.000	0.792±0.002	0.754±0.002	0.814±0.001	0.737±0.008
	2	**0.823±0.001**	0.700±0.013	0.801±0.001	0.787±0.027	0.738±0.004	0.811±0.001	0.725±0.014
	3	**0.817±0.001**	0.686±0.016	0.802±0.000	0.758±0.001	0.719±0.003	0.808±0.001	0.715±0.011
Coverage↓	1	**146.557±0.637**	235.846±4.019	160.751±0.402	154.262±1.074	167.82±0.514	**145.111±0.519**	226.33±0.594
	2	151.064±0.58	239.031±3.462	162.523±0.167	156.474±1.066	201.014±1.116	**148.997±0.233**	243.003±0.451
	3	154.721±0.646	244.44±4.735	163.031±0.364	166.064±0.535	214.51±2.727	**151.519±0.408**	245.394±0.446

由于 MLML 和 MMIB 在训练过程中综合利用了有标记样本和未标记样本,它们在多个数据集上的结果都显著优于 MLR- GL,这表明利用未标记样本能够提升弱标记分类的精度.MMIB 充分考虑了标记不平衡问题,它在多个评价度量上均取得比 MLML 更好的结果.但 MMIB 在 ESPGame,Eurlexh,Arts,Core15k 和 IAPRTC-12 数据集上的结果均被 EnWL 显著性超过.主要原因是 EnWL 在 APC 获取的特征子集上进行训练,它能够在一定程度上避免噪声和冗余特征的影响.ProDM 和 EnWL 都假设样本的标记依赖于特征信息,但 EnWL 在多个度量上都显著优于 ProDM.原因是 ProDM 最大化的是标记与高维样本的所有特征间的依赖关系,噪声和冗余特征干扰了这种依赖性度量,而 EnWL 最大化标记与选择的特征子集间的依赖.这些结果表明:相对于原始特征集合,EnWL 选择的特征子集能够更好地刻画特征和标记间的依赖关系.

EnWL,Lift 和 Tram 都是基于特征选择或维数约减的多标记学习方法,但 EnWL 在绝大多数情况下都显著优于 Tram 和 Lift.主要原因是 Tram 和 Lift 都假设已标注样本的标记信息是完整的,忽视了这些已标注样本标记缺失的特点;另一个原因是 Lift 和 Tram 分别在原始空间选择特征或降维,它们并没有显式地利用标记与特征之间的依赖关系.上述对比结果表明了本文 EnWL 的有效性.

3.4 成分分析

为了进一步调研分析 EnWL 中依赖最大化的贡献、APC 获取的特征子集的有效性以及这些特征子集上分类器集成的有效性,本文引入 3 个 EnWL 的变种:EnWL_nD,EnWL_rs 和 EnWL_nE.EnWL_nD 不考虑依赖最大化,即公式(11)中的β=0;EnWL_rs 在 T 个随机子空间[47]分别利用公式(11)训练分类器,再投票集成;特别地,随机子空间大小从{50,100,150,...,450,500}中选择.EnWL_nE 仅在 1 个 APC 获取的特征子集上训练单个分类器.与前面的实验设置类似,本部分实验中,训练集中每个有标记样本被随机隐藏 m=1 个标记,再基于这些弱标记训练样本和无标记的训练样本预测测试集中样本的标记.图 1 给出了这 4 个方法在 ESPGame 和 Arts 上的实验结果.

从这些结果可以看出,EnWL 在绝大多数情况都能获得较 EnWL_nE,EnWL_nD 和 EnWL_rs 更好的结果.具体而言,在 ESPGame 数据集上的 4 个评价度量上(除了 AvgPrec 外),这 4 种方法的性能排序(基于 Friedman 检验)为 EnWL>EnWL_nD>EnWL_rs>EnWL_nE;在 Arts 数据集上的 5 种不同评价度量中,这 4 种方法的性能排序为 EnWL>EnWL_nD>EnWL_nE>EnWL_rs.特别地,在 ESPGame 和 Arts 数据集上,EnWL_nE 和 EnWL_nD 在 5 个评价度量上的结果均被 EnWL 显著性超过,这不仅表明 EnWL 可以有效地整合多个特征子集训练的基础半监督弱标记分类器,还表明 EnWL 最大化标记与特征之间的依赖是有效的.EnWL_rs 和 EnWL 均实现了最大化样本标记和特征之间的依赖性,但 EnWL_rs 在很多评价度量上的均被 EnWL 超越,特别是在 1−HammLoss,1−RankLoss,Coverage 和 AUC 上.主要原因是 EnWL_rs 随机选择特征子集,EnWL 利用 APC 选择特征子集,这说明本文选用的 APC 可以获取具有较好表示能力的特征子集,降低噪声和冗余特征的干扰.以上对比结果分别证

明了 EnWL 在多个 APC 选择的特征子集上分别训练基于标记与特征子集之间依赖最大化的半监督弱标记分类器和集成这些基础分类器的有效性.

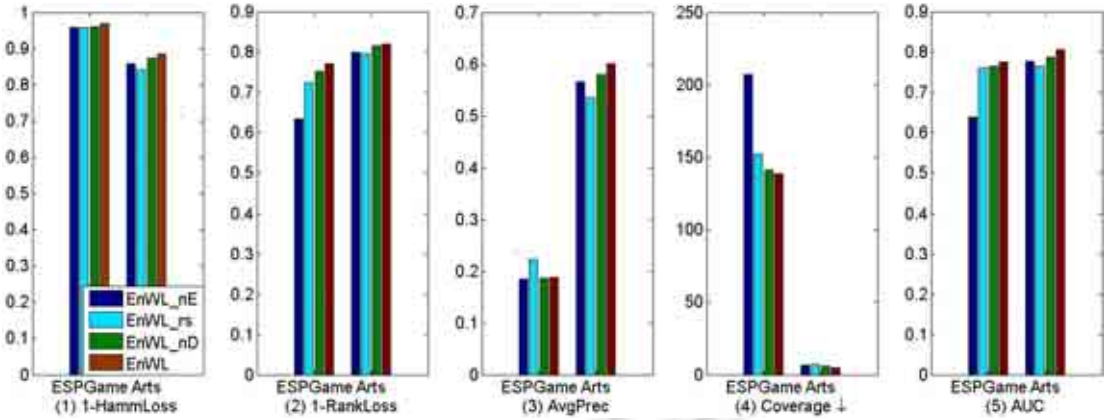

Fig.1 Experimental results of EnWL and the variants of EnWL on ESPGame and Arts

图 1 ESPGame 和 Arts 上本文 EnWL 方法及其变种的实验结果

值得注意的是,EnWL_rs 在随机选择的 T 个特征子集上分别训练分类器,而 EnWL_nE 仅在 1 个 APC 获取的特征子集训练分类器.EnWL_nE 在 Arts 数据集上均取得了比 EnWL_rs 更好的结果,这表明本文选用的 APC 可以获得具有较好表示能力的特征子集,降低噪声特征的干扰.但 EnWL_nE 在 ESPGame 上不及 EnWL_rs,原因是 ESPGame 的标记空间更大,后者集成整合了多个随机子空间上的基础分类器,在一定程度上进一步利用了标记间的关联关系.尽管 EnWL 在 EPSGame 上的 AvgPrec 略低于 EnWL_rs,但 EnWL 在其他度量上均优于后者,这些结果进一步证明了 EnWL 利用 APC 选择代表性特征子集的有效性.

3.5 基础分类器个数敏感性分析

为了分析不同数量(T)的基础分类器对 EnWL 预测性能的影响.本文统计了 EnWL 在 T 从 1 增加到 25 时对应的实验结果.与前面的实验设置类似,对训练集中的有标记样本随机隐藏 m=1 个标记构造由弱标记样本和无标记样本组成的训练集.图 2 列出了 EnWL 方法在 1−HammLoss,1−RankLoss,AvgPrec,Coverage 和 AUC 上的实验结果.

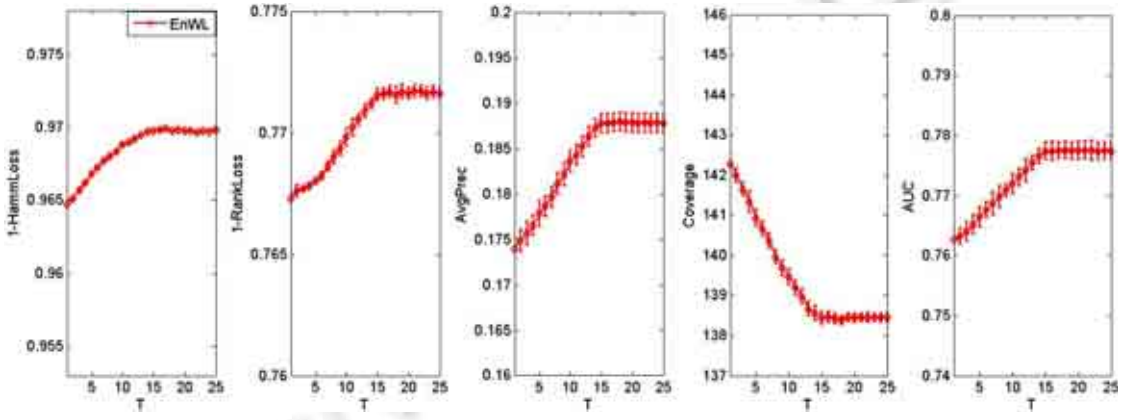

Fig.2 Experimental results of EnWL with respect to different number of base classifiers on ESPGame

图 2 ESPGame 上本文 EnWL 方法在不同基础分类器个数上的实验结果

从图 2 可以观察到:总体而言,当基础分类器个数小于 15 时,随着基础分类器个数的增加,EnWL 方法在多个评价度量上的结果都有上升;当基分类器个数大于 15 时,EnWL 方法在 5 个评价度量上的结果渐渐趋于稳定.上述实验结果表明,EnWL 易于选择合适且鲁棒的输入参数 T.

3.6 参数敏感性分析

为了分析参数α和β对 EnWL 预测性能的影响.本文统计了 EnWL 在α和β在{0.005,0.01,0.1,0.2,…,1}这 12 种取值组合情况下对应的实验结果.与前面的实验设置类似,对训练集中的有标记样本随机隐藏 m=1 个标记构造由弱标记样本和无标记样本组成的训练集.图 3 汇报了 EnWL 方法在 Core15k 和 IAPRTC-12 数据集上 1−HammLoss 的实验结果.从图 3 可以观察到:EnWL 方法在不同参数取值下的结果都较为稳定,当$\beta\in$[0.01,0.1],$\alpha\in$[0.7,0.9]时,EnWL 方法在两个数据集上通常都能取得比较稳定的较优结果.以 IAPRTC-12 数据集上的结果为例,EnWL 方法在 144 种不同参数组合情况下的实验结果都显著性优于(或等于)其他的对比算法.当β<0.01 时,EnWL 的预测性能开始下降,这表明 EnWL 最大化样本标记与特征之间的依赖可以提升弱标记学习的性能,还进一步证明结合依赖最大化的合理性.

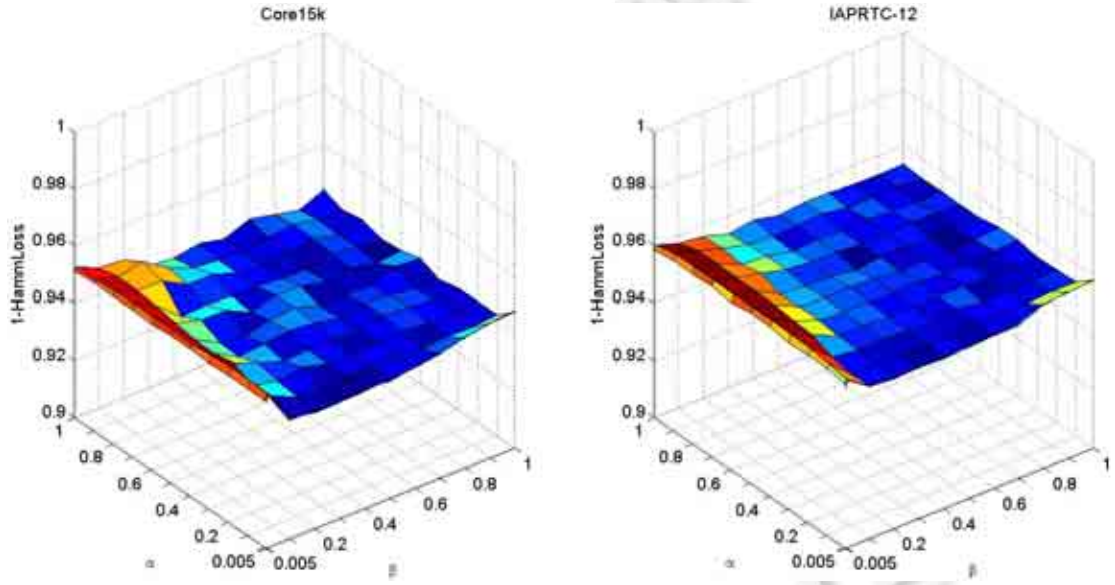

Fig.3 Performance of EnWL on Core15k and IAPRTC-12 under different combinations of α and β

图 3 在不同的α和β参数组合下本文 EnWL 方法在 Core15k 和 IAPRTC-12 上的结果

3.7 算法运行时间分析

为了统计分析各种对比算法的效率,与第 3.4 节的实验设置类似,本文统计了每种算法在相同的实验平台(CPU i5-4590,16GB RAM,Win7,Matlab2013a)下不同数据集上的运行时间,并将每个算法 5 次独立运行的平均时间报告在表 8 中.

Tabel 8 Runtime cost of comparing methods (s)

表 8 对比算法运行时间 (s)

Dataset	EnWL	MLR-GL	ProDM	MLML	MMIB	Tram	Lift
Core15k	23.83	19.99	148.34	104.84	7 296.57	99.77	363.83
ESPGame	148.39	151.66	2 276.12	2 010.85	13 479.49	1 913.24	9 565.16
Eurlex	293.09	210.59	12 561.22	11 138.10	25 479.22	10 243.77	23 188.76
IAPRTC-12	464.58	299.28	14 697.44	15 632.46	28 462.72	12 873.86	26 433.26
Arts	3 530.50	68.98	6 296.79	5 764.38	12 782.58	5 238.92	4 679.53
Reference	4 458.42	86.28	11 424.89	9 789.79	13 294.03	8 946.48	4 476.05

由表 8 可知,除 MLR-GL 外,EnWL 的运行时间总是远小于其他相关对比算法.MLR-GL 是一种监督式弱标记学习方法,它仅利用了 30%的有标记样本,且利用了优化的组稀疏和排序损失求解方法,所以其运行时间较小.ProDM,MLML 和 MMIB 由于没有对高维数据进行特别处理,直接在原始高维特征空间计算样本间相似度、训练和预测,所以其时间耗费均大于 EnWL.Lift 需要利用 *k*-means 聚类方法对每个标记寻找关键特征,所以其时间耗费也比 EnWL 大,特别是在样本和标记空间比较大的 ESPGame,Eurlex 和 IAPRTC-12 上.Tram 通过利用 MDDM 算法对高维数据进行降维,其时间耗费比 ProDM,MLML,MMIB 和 Lift 都小,但都大于 EnWL 和 MLR-GL.上述实验结果表明,本文提出的 EnWL 算法不仅能够获得比其他相关算法更高的预测精度,也获得了较这些算法(除 MLR-GL 外)更高的效率.

4 结束语

弱标记学习已被广泛应用于图像标注、文本分类和蛋白质功能预测等领域.高维数据由于特征复杂且数据量较大,更容易拥有多个语义标记和出现标记缺失.本文针对高维多标记数据提出了一种基于标记与特征依赖最大化的弱标记集成分类方法 EnWL.实验结果表明,EnWL 能够获得比其他相关算法更高的精度.由于弱标记学习的标记空间比多标记学习的标记空间更为稀疏,本文采用余弦相似性度量的可靠性有限,未来将考虑如何在稀疏的标记空间准确地刻画标记相关性和设计新的弱标记分类器集成策略,进一步提高弱标记分类精度.

References:

[1] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Trans. on Knowledge and Data Engineering, 2014,26(8): 1819−1837. [doi: 10.1109/TKDE.2013.39]

[2] Jiang Y, She QQ, Li M, Zhou ZH. A transductive multi label text categorization approach. Journal of Computer Research and Development, 2008,45(11):1817−1823 (in Chinese with English abstract).

[3] McCallum A. Multi-Label text classification with a mixture model trained by EM. In: Proc. of the Working Notes of the AAAI'99 Workshop on Text Learning. 1999. 1−7.

[4] Boutell MR, Luo J, Shen X, Brown C. Learning multi-label scene classification. Pattern Recognition, 2004,37(9):1757−1771. [doi: 10.1016/j.patcog.2004.03.009]

[5] Zhang ML, Zhou ZH. Multi-Label neural networks with applications to functional genomics and text categorization. IEEE Trans. on Knowledge and Data Engineering, 2006,18(10):1338−1351. [doi: 10.1109/TKDE.2006.162]

[6] Yu GX, Domeniconi C, Rangwala H, Zhang GJ, Yu ZW. Transductive multi-label ensemble classification for protein function prediction. In: Proc. of the 18th ACM SIGKDD on Knowledge Discovery and Data Mining. 2012. 1077−1085. [doi: 10.1145/2339530.2339700]

[7] Wang SB, Li YF. Classifier circle method for multi-label learning. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):2811−819 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4908.htm [doi: 10.13328/j.cnki.jos.004908]

[8] Li YF, Huang SJ, Zhou ZH. Regularized semi-supervised multi-label learning. Journal of Computer Research and Development, 2012,49(6):1272−1278 (in Chinese with English abstract).

[9] Kong XN, Li M, Jiang Y, Zhou ZH. A transductive multi-label classification method for weak labeling. Journal of Computer Research and Development, 2010,47(8):1392−1399 (in Chinese with English abstract).

[10] Sun YY, Zhang Y, Zhou ZH. Multi-Label learning with weak label. In: Proc. of the 25th AAAI Conf. on Artificial Intelligence. AAAI Press, 2010. 293−298.

[11] Wu BY, Liu ZL, Wang SF, Hu BG, Ji Q. Multi-Label learning with missing labels. In: Proc. of the 22nd Int'l Conf. on Pattern Recognition. 2014. 1964−1968. [doi 10.1109/ICPR.2014.343]

[12] Yu HF, Jain P, Kar P, Dhillon IS. Large-Scale multi-label learning with missing labels. In: Proc. of the 31st Int'l Conf. on Machine Learning. 2014. 593−601.

[13] Li X, Zhao FP, Guo YH. Conditional restricted Boltzmann machines for multi-label learning with incomplete labels. In: Proc. of the 16th Int'l Conf. on Artificial Intelligence and Statistics. 2015. 635−643.

[14] Wang QF, Shen B, Wang SM, Li L, Si L. Binary codes embedding for fast image tagging with incomplete labels. In: Proc. of the 13th European Conf. on Computer Vision. 2014. 425−439. [doi: 10.1007/978-3-319-10605-2_28]

[15] Yang H, Zhou JT, Cai J. Improving multi-label learning with missing labels by structured semantic correlations. In: Proc. of the 14th European Conf. on Computer Vision. 2016. 835−851. [doi: 10.1007/978-3-319-46448-0_50]

[16] Parsons L, Haque E, Lui H. Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter, 2004,6(1):90−105. [doi: 10.1145/1007730.1007731]

[17] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976. [doi: 10.1126/science.1136800]

[18] Wu BY, Lyu SW, Hu BG, Ji Q. Multi-Label learning with missing labels for image annotation and facial action unit recognition. Pattern Recognition, 2015,48(7):2279−2289. [doi: 10.1016/j.patcog.2015.01.022]

[19] Bucak SS, Jin R, Jain AK. Multi-Label learning with incomplete class assignments. In: Proc. of the 24th IEEE Conf. on Computer Vision and Pattern Recognition. 2011. 2801−2808. [doi: 10.1109/CVPR.2011.5995734]

[20] Zhao FP, Guo YH. Semi-Supervised multi-label learning with incomplete labels. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015. 4062−4068.

[21] Wu BY, Lyu SW, Ghanem B. Multi-Label learning with missing labels using a mixed graph. In: Proc. of the 25th IEEE Int'l Conf. on Computer Vision. 2015. 4157−4165. [doi: 10.1109/ICCV.2015.473]

[22] Yu GX, Domeniconi C, Rangwala H, Zhang GJ. Protein function prediction using dependence maximization. In: Proc. of the 24th European Conf. on Machine Learning. 2013. 574−589. [doi: 10.1007/978-3-642-40988-2_37]

[23] Wu BY, Lyu SW, Ghanem B. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016. 2229−2236.

[24] Zhang ML, Li YK, Liu XY. Towards class-imbalance aware multi-label learning. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015. 4041−4047.

[25] Kong XN, Ng MK, Zhou ZH. Transductive multilabel learning via label set propagation. IEEE Trans. on Knowledge and Data Engineering, 2013,25(3):704−719. [doi: 10.1109/TKDE.2011.141]

[26] Zhang Y, Zhou ZH. Multilabel dimensionality reduction via dependence maximization. ACM Trans. on Knowledge Discovery from Data, 2010,4(3):1−21. [doi: 10.1145/1839490.1839495]

[27] Zhang ML, Wu L. Multi-Label learning with label-specific features. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,31(1):107−120. [doi: 10.1109/TPAMI.2014.2339815]

[28] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003,3(5):1157−1182.

[29] Dash M, Liu H. Feature selection for clustering. In: Proc. of the 4th Pacific Asia Conf. on Knowledge Discovery and Data Mining. 2000. 98−109.

[30] Xiao Y, Yu J. Semi-Supervised clustering based on affinity propagation algorithm. Ruan Jian Xue Bao/Journal of Software, 2008, 19(11):2803−2813 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

[31] Ghamrawi N, Mccallum A. Collective multi-label classification. In: Proc. of the 14th ACM CIKM Int'l Conf. on Information and Knowledge Management. 2005. 195−200. [doi: 10.1145/1099554.1099591]

[32] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2009. 254−269. [doi: 10.1007/978-3-642-04174-7_17]

[33] Zhu XJ. Semi-Supervised learning literature. Technical Report, 1530, Department of Computer Sceinces, University of Wisconsin-Madison, 2008.

[34] Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. In: Proc. of the Advance in Neural Information Processing Systems. 2003. 321−328.

[35] Chung FRK. Spectral graph theory. In: Proc. of the Regional Conf. Series in Mathematics. 1997. 1−21.

[36] Song L, Smola A, Gretton A, Bedo J, Borgwardt K. Feature selection via dependence maximization. Journal of Machine Learning Research, 2012,13(5):1393−1434.

[37] Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Proc. of the 16th Int'l Conf. on Algorithmic Learning Theory. 2005. 63−77. [doi: 10.1007/11564089_7]

[38] Mazandu GK, CHimusa ER, Mulder NJ. Gene ontology semantic similarity: Survey on features and challenges for biological knowledge discovery. In: Proc. of the Briefings in Bioinformatics. 2016. 1−16.

[39] Tan QY, Liu YZ, Chen X, Yu GX. Multi-Label classification based on low rank representation for image annotation. Remote Sensing, 2017,9(2):No.109. [doi: 10.3390/rs9020109]

[40] Yu GX, Fu GY, Wang J, Zhu HL. Protein function prediction via semantic integration of multiple networks. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2016,13(2):220−232. [doi: 10.1109/TCBB.2015.2459713]

[41] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 2003,51(2):181−207. [doi: 10.1023/A:1022859003006]

[42] Ueda N, Saito K. Parametric mixture models for multi-labeled text. In: Proc. of the Advances in Neural Information Processing Systems. 2003. 737−744.

[43] Mencia EL, Fürnkranz J. Efficient pairwise multilabel classification for large-scale problems in the legal domain In: Proc. of the 8th European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2008. 50−65. [doi: 10.1007/978-3-540-87481-2_4]

[44] Von Ahn L, Dabbish L. Labeling images with a computer game. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. 2004. 319−326. [doi: 10.1145/985692.985733]

[45] Duygulu P, Barnard K, de Freitas JF, Forsyth DA. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. of the 7th European Conf. on Computer Vision. 2002. 97−112. [doi: 10.1007/3-540-47979-1_7]

[46] Grubinger M, Clough P, Müller H. The IAPR tc-12 Benchmark: A new evaluation resource for visual information systems. In: Proc. of the Int'l Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval. 2006. 10−20.

[47] Ho TK. The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(8):832−844. [doi: 10.1109/34.709601]

附中文参考文献:

[2] 姜远,佘俏俏,黎铭,周志华.一种直推式多标记文档分类方法.计算机研究与发展,2008,45(11):1817−1823.

[7] 王少博,李宇峰.用于多标记学习的分类圈方法.软件学报,2015,26(11):2811−2819. http://www.jos.org.cn/1000-9825/4908.htm [doi: 10.13328/j.cnki.jos.004908]

[8] 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算法.计算机研究与发展,2012,49(6):1272−1278.

[9] 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.

[30] 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报,2008,19(11):2803−2813. http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

谭桥宇(1995－),男,重庆人,硕士生,CCF 学生会员,主要研究领域为机器学习,数据挖掘.

王峻(1983－),女,博士,副教授,CCF 高级会员,主要研究领域为机器学习,数据挖掘,生物信息学.

余国先(1985－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,数据挖掘,生物信息学.

郭茂祖(1966－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为人工智能及其应用,生物信息学,近似算法,随机算法.