软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(4):1732−1748 [doi: 10.13328/j.cnki.jos.006703]

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

利用标签相关性先验的弱监督多标签学习方法*

欧阳宵[1], 陶 红[1], 范瑞东[1], 矫媛媛[2], 侯臣平[1]

[1](国防科技大学 文理学院, 湖南 长沙 410073)

[2](国防科技大学 系统工程学院, 湖南 长沙 410073)

通信作者: 侯臣平, E-mail: houchenping@nudt.edu.cn; 陶红, Email: taohong.nudt@hotmail.com

摘 要: 多标签学习是一种非常重要的机器学习范式. 传统的多标签学习方法是在监督或半监督的情况下设计的. 通常情况下, 它们需要对所有或部分数据进行准确的属于多个类别的标注. 在许多实际应用中, 拥有大量标注的标签信息往往难以获取, 限制了多标签学习的推广和应用. 与之相比, 标签相关性作为一种常见的弱监督信息, 它对标注信息的要求较低. 如何利用标签相关性进行多标签学习, 是一个重要但未研究的问题. 提出了一种利用标签相关性作为先验的弱监督多标签学习方法(WSMLLC). 该模型利用标签相关性对样本相似性进行了重述,能够有效地获取标签指示矩阵; 同时, 利用先验信息对数据的投影矩阵进行约束, 并引入回归项对指示矩阵进行修正. 与现有方法相比, WSMLLC 模型的突出优势在于: 仅提供标签相关性先验, 就可以实现多标签样本的标签指派任务. 在多个公开数据集上进行实验验证, 实验结果表明: 在标签矩阵完全缺失的情况下, WSMLLC 与当前先进的多标签学习方法相比具有明显优势.

关键词: 多标签学习; 弱监督学习; 标签相关性; 先验信息; 完全缺失标签

中图法分类号: TP18

中文引用格式: 欧阳宵, 陶红, 范瑞东, 矫媛媛, 侯臣平. 利用标签相关性先验的弱监督多标签学习方法. 软件学报, 2023, 34(4): 1732–1748. http://www.jos.org.cn/1000-9825/6703.htm

英文引用格式: Ouyang X, Tao H, Fan RD, Jiao YY, Hou CP. Weakly Supervised Multi-label Learning Using Prior Label Correlation Information. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1732−1748 (in Chinese). http://www.jos.org.cn/1000-9825/6703.htm

Weakly Supervised Multi-label Learning Using Prior Label Correlation Information

OUYANG Xiao[1], TAO Hong[1], FAN Rui-Dong[1], JIAO Yuan-Yuan[2], HOU Chen-Ping[1]

[1](College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410073, China)

[2](College of Systems Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: Multi-label learning is a very important machine learning paradigm. Traditional multi-label learning methods are designed in supervised or semi-supervised manner. Generally, they require accurate labeling of all or partial data into multiple categories. In many practical applications, it is difficult to obtain the label information with a large number of labels, which greatly restricts the promotion and application of multi-label learning. In contrast, label correlation, as a common weak supervision information, has lower requirements for labeling information. How to use label correlation for multi-label learning is an important but unstudied problem. This study proposes a method named weakly supervised multi-label learning using prior label correlation information (WSMLLC). This model restates the sample similarity by using label correlation, and can obtain label indicator matrix effectively, constrain the projection matrix of data by using prior information, and modify the indicator matrix by introducing regression terms. Compared with the existing methods, the outstanding advantage of WSMLLC model is that it can realize the label assignment of multi-label samples only by providing label correlation priors. Experimental results show that WSMLLC has obvious advantages over current advanced multi-label learning methods in the case of complete loss of label matrix.

* 基金项目: 国家自然科学基金(61922087, 61906201, 62006238, 62136005); 湖南省杰出青年基金(2019JJ20020)
收稿时间: 2021-10-21; 修改时间: 2022-04-01; 采用时间: 2022-05-06; jos 在线出版时间: 2022-07-22

Key words: multi-label learning; weakly supervised learning; label correlation; prior information; completely missing labels

多标签学习[1]是指通过一定的方法为每个样本找到与其相关的所有标签. 与传统的单标签学习相比, 多标签学习中一个样本可以同时对应多个标签. 近年来, 多标签学习问题在各个领域受到了广泛关注, 例如: 在生物信息学领域, 一个基因可能与多种个体功能有关[2]; 在文本分类中, 一个文本可能同时关联多个目标主题[3]; 在图像识别领域, 一幅图片可能包含多个待识别对象[4]; 在网络推荐系统中, 一个对象客户往往有多种商品需求[5]. 为了探寻这些问题的解决方案, 多标签学习逐渐成为各个领域专家学者关注的重点.

按训练数据中标签矩阵的完整程度, 传统的多标签分类方法可分为监督方法和半监督方法这两类. 监督的多标签分类方法又可分为问题转换类方法和算法适应类方法. 问题转换类方法是将多标签问题改造成为其他熟悉的学习场景进行解决, 如: Shan 等人[6]将多标签训练集中每一个不同的标签组合视为一个单标签分类任务的不同类别, 并提出一种共同学习的二元分类器来解决多标签分类问题; Tsoumakas 等人[7]提出了一种基于标签空间随机投影的多标签分类集成方法, 将多标签分类问题转换为多类分类问题进行解决. 算法适应类方法是对经典的学习技术进行改造, 使其能够直接处理多标签数据, 以此达到解决多标签学习问题的目的. 如 MLKNN[8]是由 k 近邻算法(KNN)发展而来, 采用最大后验概率原则给待测样本分配标签集. 注意到完全监督的多标签分类方法需要利用大量已标注的训练样本来训练, 半监督多标签分类方法将未标注数据引入训练过程中, 以缓解对大量标注样本的需求. 例如, CNMF[9]针对训练数据规模较小但标签数目较多的场景, 提出了一个新的框架来解决此类问题. CNMF 根据最小化相关性差异的原则, 采用基于样本的相关性和基于样本类成员的相关性这两种不同的相关性, 将优化问题转化为一个受约束的非负矩阵分解问题, 并提出了有效的求解算法, 该方法在文本分类任务中优势较为明显.

上述监督或半监督的多标签分类方法均需要训练数据中有已经标注的样本, 差异只是数量不同而已. 然而, 多标签学习的一大难点就在于标签信息的获取, 这一困难源于多标签学习的标记难度远高于单标签学习. 众所周知, 数据标注工作十分繁杂, 在同等研究背景下, 获取多标签样本的完整标签的代价更高、困难更大. 因此, 为一个数据集提供完整的多标签矩阵需要耗费大量的人力物力. 在大数据时代, 往往需要获取具有更大标签数量的样本以保证算法性能, 数据标注任务更为困难, 因而, 多标签样本标注工作的高成本高代价的特性一直制约着多标签学习的发展.

相对于要求严苛的多标签标注而言, 现实生活中存在许多弱标记信息, 如标签完整度信息、标签稀疏性和标签相关性等. 标签完整度信息是指与数据集对应的标签矩阵信息是否完整以及标签矩阵的信息含量, 如 Huang 等人[10]提出的 MCUL 探讨了部分标签值缺失和部分标签完全缺失时的多标签学习问题, 利用标签相关性解决了标签缺失问题. 标签稀疏性是指某一标签占比太大, 常指不平衡类的学习问题. 如 Luo 等人[11]研究了样本失衡的问题, 引入非对称分段损失函数, 通过加大正负样本分类边界并修改损失函数类型, 提高少数标签的识别精度. 标签相关性则指标签间的关联信息, 常包含一阶策略、二阶策略和高阶策略[1]. 二元关联是一种典型的一阶策略, Zhang 等人[12]基于标签间的二元关联对多标签学习方法展开了详尽的论述. 二阶策略通常考虑标签间的成对关系, 如 Weng 等人[13]利用局部成对标签相关性, 通过整合其他标签特异性特征来扩展每个标签的特异性特征. HOT-VAE[14]是一个典型的基于高阶标签相关性进行自适应学习的例子, 它不仅可以获得多个对象的内在关系, 而且此关系对于任何特征的变化具有适应性.

在各种弱标记信息中, 标签相关性是一种重要且常见的先验信息, 所谓标签相关性先验, 是指相同标签组合之间的相关性在不同数据集中应该具有一致性. 例如, 凭借先验信息, 我们可以知道在一般的自然景象图片中, 山脉和树木同时出现的可能性很大, 而沙漠、树木和日落同时出现的可能性则很小, 那么给定一张自然风景图片, 我们就可以利用先验信息判断出它同时包含山脉和树木的可能性远大于同时包含树木、沙漠和日落的可能性. 如图 1 所示, 本文统计了数据集 Espgame 和 Pascal07 (http://lear.inrialpes.fr/people/guillaumin/data.php)的标注情况, 它们共有标签 Bird, Car 和 Dog 的相关性在两个不同数据集中是相似的, 就可以作为多标签学习的先验输入. 如何利用此种广泛存在且容易获取的标签相关性信息来解决没有标注训练数据的多标

签学习问题, 是一个重要且尚未研究的课题.

针对上述问题, 本文提出了一个全新的框架以解决无标记样本的多标签聚类问题, 即利用标签相关性先验解决无标注矩阵的多标签数据集的标签指派问题. 如图 1 所示, 利用标签相关性在不同数据集中的一致性得到标签相关性矩阵后, 将其作为先验输入, 同时结合对样本相似性的学习和对指示矩阵的回归, 可以得到多标签预测结果. 其中, D, C, N 分别表示数据集的维数、标签数目和样本个数; 黑白颜色展示的是样本对应的标签矩阵, 例如第 i 行第 j 个方框为黑色, 表示第 i 个标签属于第 j 个样本, 否则不属于. 因此, 本文以标签相关性矩阵作为权重矩阵对样本相似性进行描述, 通过从不同角度对相似性的描述, 可导出标签指示矩阵. 利用先验信息对数据的投影矩阵进行限制, 然后结合回归思想对标签指示矩阵进行修正, 实现了对无标记矩阵数据集的多标签指派问题.

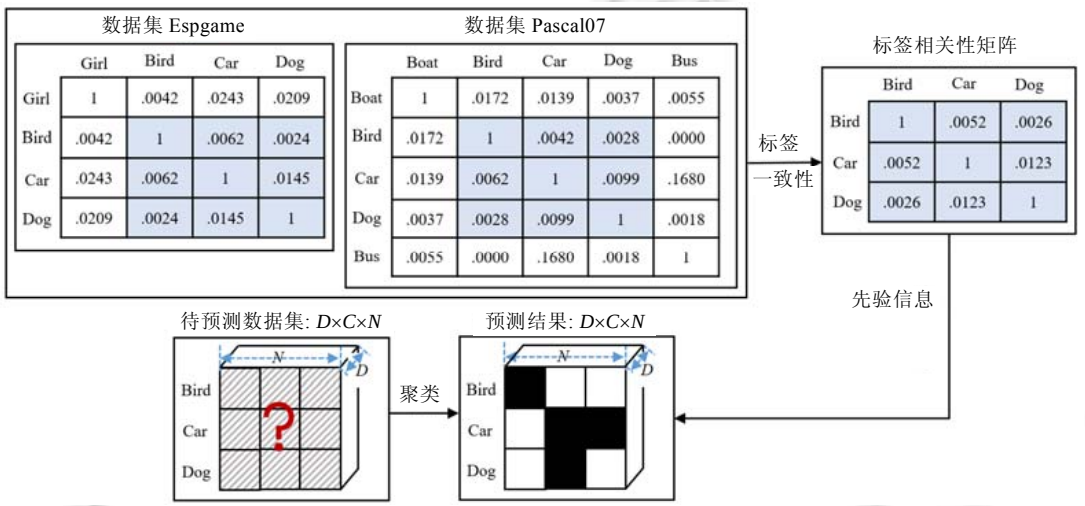


图 1 利用标签相关性的弱监督多标签学习示意图

本文的主要贡献总结如下:

(1) 在多标签学习任务中, 以往的方法都需要利用多标签的标注矩阵, 此标注矩阵可能是完全的、也可能是部分缺失或引入了噪声项的; 然后, 利用标注信息建立分类模型, 但此类模型只适用于具有标签矩阵的多标签数据集. 当多标签数据集的标签矩阵难以获取时, WSMLLC 模型仅需提供历史标签相关性矩阵, 就可以实现对多标签数据集的标签指派目标;

(2) 本文采用的标签相关性信息常见且易于获取, 此外, WSMLLC 模型采用两种不同的方法从不同的角度描述样本相似性, 通过使用带有指示矩阵的回归项来指导多标签聚类模型. WSMLLC 模型可以完成多标签的聚类任务, 因此适用于各种类型的多标签数据集;

(3) 本文利用标签相关性先验完成了聚类的任务, 避免了无监督多标签聚类中标签难以对齐的情形(C 个标签对齐方式有 2^C 种), 减少了计算量;

(4) 本文与当前具有代表性的有监督多标签分类方法及半监督多标签分类方法进行比较, 通过设计多组对比实验及收敛性理论分析, 对算法的性能进行了全面评估. 实验结果表明, 本文提出的方法显著优于其他同类方法.

本文第 1 节介绍弱监督多标签学习的相关工作. 第 2 节详细介绍本文提出的理论方法及求解思路, 并对目标函数的收敛性进行证明. 第 3 节采用多个对比实验, 验证本文所提模型的有效性. 第 4 节总结全文工作, 并讨论弱监督多标签学习的挑战及未来的发展趋势.

1 相关工作

随着社会需求的发展, 各种情形的弱监督多标签学习方法被不断提出. 一般而言, 弱监督方法可分为不准确监督、不确切监督和不完全监督这 3 种类型[15], 因此, 弱监督多标签学习也可以按照此种方式进行划分. 现将弱监督多标签学习的相关工作总结如下.

- 多标签的不准确监督是指有关标签矩阵的信息并不总是准确无误的, 常指标签信息质量差, 如标签损坏、标记不准确等, 其中一个典型的情形就是多标签中混入了噪声等干扰因素. 由于不准确的标签矩阵会降低模型的预测能力, 因此, 多标签不准确监督的关键是克服有噪声标签所带来的影响, 并生成一个鲁棒的预测模型来估计测试集样本的标签矩阵. 为此, Frenay 等人[16]对标签噪声进行了全面的阐述, 考虑了标签噪声的定义和来源、类型及分类, 讨论了标签噪声的潜在后果, 并综合阐述了标签噪声鲁棒、净化和容忍算法等. Sun 等人[17]提出了一种基于低秩和稀疏约束的成本敏感标签排序方法, 不仅可以去除噪声标签, 同时还能恢复缺失标签. 尽管 CORALS 假设了噪声标签是稀疏的, 但它充分挖掘了不同标签之间的相关性;
- 多标签的不确切监督是指提供的标签矩阵与实际预期有差别, 一个典型的场景就是只提供关于多标签的粗粒度的标记, 而不是对数据的精确标记, 其目的是利用粗粒度的训练数据来预测未知数据的准确标签矩阵. 解决此问题的主要方法之一即为多标签多示例学习. 目前, 多标签多示例学习已广泛应用于图像注释[18]、视频标注[19,20]、基因检测[21]等领域中. 如视频标签一般无法覆盖视频所有内容, Zeng 等人[19]提出了一个多标签多示例学习框架, 从视频级别的标签中学习镜头标签, 降低标签成本, 同时还可以通过控制视频标签和镜头之间的关系来学习标签语义, 解决粗粒度问题;
- 多标签的不完全监督主要是处理标记矩阵信息不全的数据集, 此时, 训练集中既有标记数据也有未标记数据, 但标记数据数量有限, 不足以通过完全监督的方式训练得到一个好的分类器, 因此, 多标签不完全监督学习的目标是获得比仅从已标记数据中学得的分类器更好得性能. 如 Huang 等人[22]探索了隐藏在数据中的潜在标签, 并可同时预测具有潜在标签和已知标签的新样本. Cheng 等人[23]研究了标签矩阵部分缺失情况下的多标签学习问题, 利用从不完全训练数据集中学习的正标签相关性和负标签相关性来补全缺失的标签, 然后通过选择标签特定特征得以实现多标签分类. Zhu 等人提出了一种新的多标签相关学习方法 GLOGAL[24], 该方法通过学习潜在标签表示和优化标签流形, 可以在训练分类器的同时探索全局和局部的标签相关性, 并恢复缺失的标签.

虽然上述 3 类弱监督多标签学习方法大大降低了对标注矩阵的要求, 但其仍需一定量的标签矩阵才能实现多标签的指派问题, 因此, 这些监督信息仍不够弱. 尽管无标注矩阵的弱监督多标签学习困难重重, 但其研究意义十分重大, 近几年, 只有极少数文章对此进行了初步探索, 如 Word2Cluster[25]采用的是类似于多类聚类的策略, 在聚类完成后合并相似的类, 达到多标签学习的目的. 但此方法主要为策略分析, 缺乏实验验证. 为进一步拓展多标签的应用场景, 本文提出了一种利用标签相关性先验的弱监督多标签学习方法, 它可以在标签矩阵完全缺失的情况下, 完成对多标签数据集的标签指派任务.

2 利用标签相关性的弱监督多标签学习模型

本节主要介绍利用先验信息的弱监督多标签学习模型, 可为无标注样本分配标签. 本节内容在没有标注矩阵的前提下, 仅利用标签相关性导出目标函数式, 并给出目标函数的优化求解思路及收敛性证明.

2.1 问题描述

本文所提的利用先验信息的弱监督多标签学习问题的研究是指在给定无标注矩阵的前提下, 仅利用从历史标签数据中获取的标签相关性信息来指导多标签的指派问题, 然后建立一个聚类模型. 表 1 展示了本文符号的含义. 一般而言, 用黑体大写斜体字母表示矩阵, 黑体小写斜体字母表示向量. 对于矩阵 $\boldsymbol{R}=(r_{ij})_{C\times C}$ 而言, $\boldsymbol{R}$ 的第 i 行第 j 列元素记为 r_{ij}, 矩阵 $\boldsymbol{R}$ 的 F 范数定义为 $\|\boldsymbol{R}\|_F^2=Tr(\boldsymbol{R}^T\boldsymbol{R})=\sum_{i,j=1}^{C}|r_{ij}|^2$, $\boldsymbol{R}$ 的第 j 列可记为 $\boldsymbol{r}_{\cdot j}$, 向

量的 2 范数定义为 $\|\boldsymbol{r}_{\cdot j}\|_2^2=\sum_{i=1}^{C}|r_{ij}|^2$, $\mathbf{1}$ 表示值全为 1 的列向量.

表 1 符号及含义

符号	含义	符号	含义	符号	含义
$\boldsymbol{X}$	特征矩阵	$\boldsymbol{S}$	样本相关性矩阵	$\boldsymbol{W}$	$\boldsymbol{X}$ 的系数矩阵
$\boldsymbol{Y}$	标签矩阵	$\boldsymbol{R}$	标签相关性矩阵	C	标签数目
N	样本数目	$\boldsymbol{T}$	标签指示矩阵	D	$\boldsymbol{X}$ 的特征维数

本文的目标就是: 在给定数据特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{N\times D}$, 标签相关性矩阵 $\boldsymbol{R}\in\mathbb{R}^{C\times C}$, 在数据的标签矩阵 $\boldsymbol{Y}\in\mathbb{R}^{N\times C}$ 不可用的前提下, 通过 WSMLLC 模型得到样本的标签属性矩阵, 即指示矩阵 $\boldsymbol{T}\in\mathbb{R}^{N\times C}$, $\boldsymbol{T}$ 中元素 $t_{ij}\geqslant 0$, 后文中用 $\boldsymbol{t}_i=[\boldsymbol{T}_{i1},\boldsymbol{T}_{i2},\ldots,\boldsymbol{T}_{ic}]^T$ 和 $\boldsymbol{t}_j=[\boldsymbol{T}_{j1},\boldsymbol{T}_{j2},\ldots,\boldsymbol{T}_{jc}]^T$ 分别表示将每个类标签分配给第 i 和第 j 个样本的置信度.

2.2 模型与方法

此项工作背后隐含着两个重要的假设, 假设 1 是: 如果历史数据与当前数据在分布上具有高度的相似性, 那么它们将分配得到一组相似的标签集合; 假设 2 是: 如果两个样本具有相似的特征输入模式, 那么它们的类标签集合往往具有较高的重叠度[9]. 基于上述假设, 本文从特征和标签这两个维度来衡量样本之间的相似程度. 具体来说, 特征维度是基于输入的特征数据, 标签维度是基于样本分配得到的标签成员, 按照计算原理, 可将其分别命名为基于特征的样本相似度和基于标签的样本相似度.

首先介绍如何计算基于特征的样本相似度. 具体来说, 对于多标签数据集中具有多个标签的样本 $\boldsymbol{x}_i$, 其能通过数据的相似性矩阵 $\boldsymbol{S}\in[0,1]^{N\times N}$ 与其他所有样本建立联系. 一般而言, 样本间的距离越近($\|x_i-x_j\|_2^2$ 越小), 对应的相似性 s_{ij} 越大, 即样本间的距离 $\|x_i-x_j\|_2^2$ 与它们之间的相似性 s_{ij} 呈负相关. 为了平衡样本距离与相似性之间的关系, 本文在目标函数中引入加权项来约束相似性矩阵 $\boldsymbol{S}$, 此时, 相似性矩阵 $\boldsymbol{S}$ 可通过问题(1)进行求解:

$$\left.\begin{aligned}&\min_{s_i}\sum_{i,j=1}^{N}\|x_i-x_j\|_2^2\,s_{ij}+\alpha\|\boldsymbol{S}\|_F^2\\&\text{s.t. }\forall i,\ s_i\mathbf{1}=1,\ 0\leqslant s_{ij}\leqslant 1\end{aligned}\right\}\tag{1}$$

其中, $\boldsymbol{s}_i$ 为相似性矩阵 $\boldsymbol{S}$ 的第 i 个向量, 它的第 j 个元素为 s_{ij}.

对于基于标签的样本相似度, 任意两个样本间的相似性可通计算它们在标签成员上的重叠度来估计. 例如: 传统方法中, 将样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似性描述为 $\boldsymbol{t}_i^T\boldsymbol{t}_j$. 然而, 这种计算方法独立地对待每个标签, 忽略了标签之间的相关关系. 举例来说, 如果两个样本没有共同标签, 经上述方法计算得到的基于标签的样本相似性必定为 0, 从而引出一个新的问题: 如果两个样本之间的相似性很强, 但却没有共同标签时该怎么处理. 例如分别来自层次分类法中子类与父类两个样本, 它们的类别标签可能没有重叠, 但样本间的相关性却很强[9]. 为了克服传统方法的缺点, 本文通过引入标签相关性矩阵来完善基于标签的样本相似度的计算. 具体而言, 采用一种加权点乘的计算方法, 即 $\boldsymbol{t}_i^T\boldsymbol{R}\boldsymbol{t}_j$, 来计算样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似性, 其中, $\boldsymbol{R}=(r_{kl})_{C\times C}$ 为标签相关性矩阵; 元素 r_{kl} 表示第 k 个标签与第 l 个标签之间的相关程度, 可以通过标签的余弦相似度来近似. $\boldsymbol{t}_i^T\boldsymbol{R}\boldsymbol{t}_j$ 中 $\boldsymbol{R}$ 也可看作是一个权重矩阵, 它根据标签之间关系的强弱对标签指示矩阵进行加权, 以此获得与实际情况差距较小的样本相似性. WSMLLC 从特征和标签这两个维度充分考虑了样本间所有可能建立的联系, 从而能够充分学得样本间的相似性关系. 由于 s_{ij} 和 $\boldsymbol{t}_i^T\boldsymbol{R}\boldsymbol{t}_j$ 均表示样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的相似性, 因此我们期望 s_{ij} 和 $\boldsymbol{t}_i^T\boldsymbol{R}\boldsymbol{t}_j$ 两者之间的差距不会太大. 从而可以归纳得到如下优化目标:

$$\left.\begin{aligned}&\min_{\mathrm{T}}\sum_{i,j=1}^{N}\left(s_{ij}-\sum_{k,l=1}^{C}t_{ik}r_{kl}t_{jl}\right)^2\\&\text{s.t. }\forall i,\ t_{ik}\geqslant 0,\ i=1,2,\ldots,N;\ k=1,2,\ldots,C\end{aligned}\right\}\tag{2}$$

其中, N 为样本数目, C 为标签数目.

综上, 本文在权衡样本距离与相似性的基础上, 同时学习基于特征的样本相似性 $\boldsymbol{S}$ 和基于标签的样本相似性 $\boldsymbol{t}_i^T \boldsymbol{R} \boldsymbol{t}_j$, 最小化它们之间的差异可以获得最接近实际情况的指示矩阵 $\boldsymbol{T}$. 结合公式(1)和公式(2)可以得到:

$$\min_{\boldsymbol{S},\boldsymbol{T}} \sum_{i,j=1}^{N} \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2^2 s_{ij} + \lambda_1 \| \boldsymbol{S} - \boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}} \|_F^2 \tag{3}$$

其中, 指示矩阵 $\boldsymbol{T} \in \mathbb{R}^{N \times C}$ 为优化的关键目标. 为了充分利用先验信息 $\boldsymbol{R}$, 提高指示矩阵的准确度, 本文还引入了回归项对 $\boldsymbol{T}$ 进行修正. $\boldsymbol{X}\boldsymbol{W}$ 可以看作是一个分类器, $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_C] \in \mathbb{R}^{D \times C}$ 为投影矩阵, 它将数据映射到我们需要学习的指示矩阵 $\boldsymbol{T}$. $\boldsymbol{W}$ 的每一列表示投影后的类别, 即 $\boldsymbol{w}_k$ 和 $\boldsymbol{w}_l$ 分别表示将数据映射到第 k 类和第 l 类. 同时, 引入基于利用标签相关性矩阵 $\boldsymbol{R}$ 的正则化项, 来使 $\boldsymbol{W}$ 更准确地反映数据矩阵与标签矩阵之间的回归关系. 具体而言, 如果标签 C_k 与标签 C_l 之间具有很强的相关性, 那么它们之间的相关系数 r_{kl} 应该较大, 且它们对应的投影向量 $\boldsymbol{w}_k$ 和 $\boldsymbol{w}_l$ 也应具有很强的相似性, 因此, $\| \boldsymbol{w}_k - \boldsymbol{w}_l \|_2^2$ 的值应该比较小, 即 $\boldsymbol{w}_k$ 和 $\boldsymbol{w}_l$ 之间的欧氏距离与标签相关系数 r_{kl} 具有负相关关系. 因此, 采用与公式(1)式类似的策略, 得到优化目标如公式(4)所示:

$$\min_{\boldsymbol{W},\boldsymbol{T}} \| \boldsymbol{X}\boldsymbol{W} - \boldsymbol{T} \|_F^2 + \sum_{k,l=1}^{C} \| \boldsymbol{w}_k - \boldsymbol{w}_l \|_2^2 r_{kl} \tag{4}$$

综合考虑公式(3)和公式(4), 本文建立了一个利用标签相关性先验的弱监督多标签学习模型:

$$\begin{aligned} &\min_{\boldsymbol{S},\boldsymbol{T},\boldsymbol{W}} \sum_{i,j=1}^{N} \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2^2 s_{ij} + \lambda_1 \| \boldsymbol{S} - \boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}} \|_F^2 + \lambda_2 \| \boldsymbol{X}\boldsymbol{W} - \boldsymbol{T} \|_F^2 + \lambda_3 \sum_{k,l=1}^{C} \| \boldsymbol{w}_k - \boldsymbol{w}_l \|_2^2 r_{kl} \\ &\text{s.t. } \forall i,j,k,\ s_{ij},t_{ik} \geqslant 0;\ \sum_{j=1}^{N} s_{ij} = 1;\ i,j = 1,2,...,N;\ k,l = 1,2,...,C \end{aligned} \tag{5}$$

其中, λ_1, λ_2 和 λ_3 均为大于 0 的参数.

最后输出的指示矩阵 $\boldsymbol{T}$, 它的每一行表示每个样本对应的标签信息, 标签排序与输入的标签相关性矩阵保持一致. 因为标签属性只有 1(属于)和−1(不属于)这两种, 因此将 $\boldsymbol{T}$ 通过归一化方法转化为软标签矩阵 $\hat{\boldsymbol{T}} = (\hat{t}_{ij})$, $\forall \hat{t}_{ij} \in [0,1]$ 后, 需要引入阈值 z 进行硬化, 才能得到最终预测的标签矩阵. z 的具体取值将在实验部分详细介绍.

2.3 模型求解

本文采用交替迭代的方式来优化公式(5)中的目标函数. 由目标函数(5)可知, 一共有 3 个变量需要优化. 在迭代过程中, 3 个变量依照次序分别优化, 即: 在当前步骤更新某个变量时, 另外两个变量是固定的, 可视为常数. 此外, 本文用大写字母 F 表示目标函数.

2.3.1 固定 $\boldsymbol{W}$ 和 $\boldsymbol{T}$, 优化 $\boldsymbol{S}$

目标函数(5)中, 最后两项可以看作是常数, 此时的优化子问题变为

$$\min_{\boldsymbol{S} \in [0,1]^{N \times N}, \boldsymbol{S}\boldsymbol{1}=\boldsymbol{1}} \sum_{i,j=1}^{N} \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2^2 s_{ij} + \lambda_1 \sum_{i,j=1}^{N} \left(s_{ij} - \sum_{k,l=1}^{C} t_{ik} r_{kl} t_{jl} \right)^2 \tag{6}$$

为简化计算, 记向量 $\boldsymbol{d}_i$ 的第 j 个元素为 $d_{ij}^x = \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2^2$, 此时, 优化问题可转换为

$$\min_{\boldsymbol{s}} \sum_{i,j=1}^{N} \left(s_{ij} + \frac{d_{ij}^x}{2\lambda_1} - [\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}}]_{ij} \right)^2 \tag{7}$$

其中, $[\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}}]_{ij}$ 表示样本 i 与样本 j 之间的相似性, 记 $h_{ij} = \frac{d_{ij}^x}{2\lambda_1} - [\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}}]_{ij}$, 结合公式(5)中的约束条件, 问题(7)可表述为

$$\left.\begin{aligned} &\min_{\boldsymbol{s}} \sum_{i,j=1}^{N} (s_{ij} + h_{ij})^2 \\ &\text{s.t. } s_i^{\mathrm{T}} \boldsymbol{1} = 1;\ 0 \leqslant s_{ij} \leqslant 1 \end{aligned}\right\} \tag{8}$$

此时, 可通过拉格朗日乘数法并结合 KKT 条件[26]解决问题(8).

最后, 为保证 $\boldsymbol{S}$ 的对称性, 需要将得到的矩阵 $\boldsymbol{S}$ 对称化, 采用的方法为 $\boldsymbol{S}=(\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}})/2$.

2.3.2 固定 $\boldsymbol{T}$ 和 $\boldsymbol{S}$, 优化 $\boldsymbol{W}$

此时, 目标函数可简化为

$$\min_{\boldsymbol{W}\in\mathbb{R}^{D\times C}} \lambda_2 \| \boldsymbol{XW}-\boldsymbol{T} \|_F^2 +\lambda_3 \sum_{k,l=1}^{C} \| \boldsymbol{w}_k-\boldsymbol{w}_l \|_2^2 r_{kl} \tag{9}$$

根据图理论, 优化目标(9)等价于优化目标(10):

$$\min_{\boldsymbol{W}\in\mathbb{R}^{D\times C}} \lambda_2 \| \boldsymbol{XW}-\boldsymbol{T} \|_F^2 +2\lambda_3 Tr(\boldsymbol{WLW}^{\mathrm{T}}) \tag{10}$$

其中, $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{R}$ 是标签相关矩阵 $\boldsymbol{R}$ 对应的图 Laplacian 矩阵, $\boldsymbol{D}=(d_{kl})_{l\times l}$ 是 $\boldsymbol{R}$ 的对角阵, 其对角元素 $d_{kk}=\sum_{l=1}^{C} r_{kl}$. 对公式(10)关于 $\boldsymbol{W}$ 求导并令结果为 0, 可以得到形如 $\boldsymbol{AX}+\boldsymbol{XB}=\boldsymbol{C}$ 的 sylvester 方程:

$$\lambda_2 \boldsymbol{X}^{\mathrm{T}}\boldsymbol{XW}+2\lambda_3 \boldsymbol{WL}=\lambda_2 \boldsymbol{X}^{\mathrm{T}}\boldsymbol{T} \tag{11}$$

然而, 当 $\boldsymbol{X}^T\boldsymbol{X}$ 与 $\boldsymbol{L}$ 有重合的特征值时, 方程(11)的解不唯一. 因此, 本文采用梯度下降的方法求解 $\boldsymbol{W}$, 方程(10)可等价于:

$$\min_{\boldsymbol{W}} Tr[\boldsymbol{W}^{\mathrm{T}}(\lambda_2 \boldsymbol{X}^T\boldsymbol{X})\boldsymbol{W}]+Tr(-2\lambda_2 \boldsymbol{T}^T\boldsymbol{XW})+Tr[\boldsymbol{W}(2\lambda_3 L)\boldsymbol{W}^{\mathrm{T}}] \tag{12}$$

假设我们完成了第 k 次迭代, 用 $\boldsymbol{W}_k$ 表示第 k 次更新后得到的 $\boldsymbol{W}$ 值, 那么可以得到公式(10)中的梯度如下:

$$\nabla F(\boldsymbol{W}_k)=2\lambda_2 \boldsymbol{X}^T\left(\boldsymbol{XW}_k-\boldsymbol{T}\right)+2\lambda_3\left(\boldsymbol{W}_k\boldsymbol{L}^T+\boldsymbol{W}_k L\right) \tag{13}$$

因此, 第 k+1 次迭代过程可表述为

$$\boldsymbol{W}_{k+1}=\boldsymbol{W}_k-\tau_k \nabla F(\boldsymbol{W}_k) \tag{14}$$

其中, τ_k 为迭代步长, 可由精确线性搜索方法[27]确定:

$$\tau_k=\arg\min_{\tau} F(\boldsymbol{W}_{\boldsymbol{k}}-\tau\nabla F(\boldsymbol{W}_{\boldsymbol{k}})) \tag{15}$$

2.3.3 固定 $\boldsymbol{S}$ 和 $\boldsymbol{W}$, 优化 $\boldsymbol{T}$

优化 $\boldsymbol{T}$ 时, 通过观察可以知道, 目标函数公式(5)中只有第 2 项、第 3 项与 $\boldsymbol{T}$ 有关, 因此只需要优化这两项即可. 由于直接优化存在一定难度, 故本文将这两部分分别放大后合并, 通过优化上界来达到优化目标函数的目的.

针对目标函数(5)中的第 2 项, 可将其按如下方式放大:

$$\begin{aligned}\sum_{i,j=1}^{N}\left(s_{ij}-\sum_{k,l=1}^{C} t_{ik}r_{kl}t_{jl}\right)^2 &\leqslant \sum_{i,j=1}^{N}\sum_{k,l=1}^{C}\frac{\tilde{t}_{ik}r_{kl}\tilde{t}_{jl}}{[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}}\left(s_{ij}-[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}\frac{t_{ik}r_{kl}t_{jl}}{[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}}\right)^2\\ &\leqslant \sum_{i,j=1}^{N}\left(s_{ij}^2+\sum_{l=1}^{C}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}\frac{t_{jl}^4}{\tilde{t}_{jl}^3}-4\sum_{l=1}^{C}s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{jl}\tilde{t}_{jl}\lg t_{jl}-2s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}+4\sum_{k=1}^{C}s_{ij}\tilde{t}_{ik}[\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{kj}\lg\tilde{t}_{ik}\right)\end{aligned} \tag{16}$$

针对目标函数(5)中的第 3 项, 经过放大可得到公式(17):

$$\sum_{i=1}^{N}\sum_{k=1}^{C}([\boldsymbol{XW}]_{ik}-t_{ik})^2\leqslant\sum_{i=1}^{N}\sum_{k=1}^{C}(t_{ik}^2-2[\boldsymbol{XW}]_{ik}\tilde{t}_{ik}(1+\lg t_{ik}-\lg\tilde{t}_{ik})+[\boldsymbol{XW}]_{ik}^2) \tag{17}$$

因此, 将目标函数(5)中的第 2 项、第 3 项替换为不等式(16)和不等式(17)的放大上界, 可得到优化 $\boldsymbol{T}$ 上界的子优化目标:

$$\begin{aligned}&\min_{\boldsymbol{S},\boldsymbol{T},\boldsymbol{W}}\sum_{i,j=1}^{N}\| \boldsymbol{x}_i-\boldsymbol{x}_j \|_2^2 s_{ij}+4\lambda_1\sum_{i,j=1}^{N}\sum_{k=1}^{C}s_{ij}\tilde{t}_{ik}[\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{kj}\lg\tilde{t}_{ik}+\\ &\lambda_1\sum_{i,j=1}^{N}\left(s_{ij}^2+\sum_{k,l=1}^{C}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}\frac{t_{jl}^4}{\tilde{t}_{jl}^3}-4\lambda_1\sum_{k,l=1}^{C}s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{jl}\tilde{t}_{jl}\lg t_{jl}-2s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}\right)+\\ &\lambda_2\sum_{i=1}^{N}\sum_{k=1}^{C}(t_{ik}^2-2[\boldsymbol{XW}]_{ik}\tilde{t}_{ik}(1+\lg t_{ik}-\lg\tilde{t}_{ik})+[\boldsymbol{XW}]_{ik}^2)+\lambda_3\sum_{k,l=1}^{N}\| \boldsymbol{w}_k-\boldsymbol{w}_l \|_2^2 r_{kl}\end{aligned} \tag{18}$$

对边界函数(18)求导并令其为 0, 得到的结果如下:

$$4\lambda_1\sum_{i=1}^{N}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{jl}\frac{t_{jl}^3}{\tilde{t}_{jl}^3}-4\lambda_1\sum_{i=1}^{N}s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}\tilde{t}_{jl}\frac{1}{t_{jl}}+\lambda_2\left(2t_{jl}-2[\boldsymbol{X}\boldsymbol{W}]_{jl}\tilde{t}_{jl}\frac{1}{t_{jl}}\right)=0 \tag{19}$$

整理并分析公式(19), 可得到目标变量 $\boldsymbol{T}$ 的闭式解(20):

$$t_{jl}=\sqrt{\frac{-\lambda_2\tilde{t}_{jl}^3+\sqrt{\lambda_2^2\tilde{t}_{jl}^6+8\lambda_1u_{jl}\tilde{t}_{jl}^4(2\lambda_1v_{jl}+\lambda_2[\boldsymbol{X}\boldsymbol{W}]_{jl})}}{4\lambda_1u_{jl}}} \tag{20}$$

其中, $u_{jl}=[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{T}\tilde{\boldsymbol{T}}\boldsymbol{R}]_{jl}$, $v_{jl}=[\boldsymbol{S}\tilde{\boldsymbol{T}}\boldsymbol{R}]_{jl}$. 此外, 解析式(20)中, $\boldsymbol{T}$ 表示当前迭代时的指示矩阵, $\tilde{\boldsymbol{T}}$ 表示的是经过上一步迭代得到的指示矩阵.

最后, 对于得到的指示矩阵 $\boldsymbol{T}$, 它的取值范围不是固定的, 为了使结果更加直观以便于理解, 本文将矩阵 $\boldsymbol{T}$ 进行归一化得到 $\hat{\boldsymbol{T}}$. $\hat{\boldsymbol{T}}$ 也称为软标签矩阵, 它的取值范围是[0,1]. 为保证最终输出的结果为每个样本的具体标签属性, 本文以 z 为阈值进行划分, 当元素 $\hat{t}_{ij}>z$ 时, 可将标签 j 分配给样本 i, 即对应的标签矩阵 $\boldsymbol{Y}_{ij}$=1; 否则认为标签 j 不属于样本 i, $\boldsymbol{Y}_{ij}$=−1.

算法 1. WSMLLC.

输入: 无标注数据矩阵 $\boldsymbol{X}$, 标签相关性矩阵 $\boldsymbol{R}$, 参数λ_1, λ_2 和λ_3;

输出: $\boldsymbol{X}$ 的软标签矩阵 $\hat{\boldsymbol{T}}$.

步骤:

1. 随机初始化矩阵 $\boldsymbol{T}$;

重复

2. 固定矩阵 $\boldsymbol{T}$ 和 $\boldsymbol{W}$, 使用公式(8)更新 $\boldsymbol{S}$;
3. 固定矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{T}$, 使用公式(12)更新 $\boldsymbol{W}$;
4. 固定矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{S}$, 使用公式(20)更新 $\boldsymbol{T}$;

直到收敛

本文中一共有 3 个变量有待优化, 我们分析了每一步迭代中更新每个变量的复杂度. 算法 1 中, $\boldsymbol{S}$ 的更新最为简单, 基本不涉及较为复杂的运算, 更新 $\boldsymbol{S}$ 的时间复杂度仅需 O(n). 步骤 3 中, $\boldsymbol{W}$ 的更新采用的是梯度下降的求解方法, 由公式(13)可知, 计算梯度 $\nabla_F(\boldsymbol{W}_k)=Tr[\boldsymbol{W}_k^{\mathrm{T}}(\lambda_2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}_k)]+Tr[-2\lambda_2\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}_k]+Tr[\boldsymbol{W}_k(2\lambda_3L)\boldsymbol{W}_k^{\mathrm{T}}]$ 的时间复杂度为 O(ND^2+D^2C+DC^2). 最后, 步骤 4 中更新 $\boldsymbol{T}$ 主要是通过控制目标函数的上界进行优化, 得到的结果为闭式解, 时间复杂度为 O(NDC). 因此, 假设算法 1 总的迭代次数为η, 那么算法 1 总的时间复杂度可表述为 O(η(ND^2+D^2C+DC^2+NDC)). 由此可知, 本文方法的计算复杂度相对于样本容量 N 是线性的, 具有很强的可扩展性, 适用于大规模数据集的多标签预测问题.

2.4 收敛性证明

定理 1. 采用算法 1 中的优化程序, WSMLLC 的目标函数值非增.

证明: 令:

$$F(\boldsymbol{S},\boldsymbol{W},\boldsymbol{T})=\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2s_{ij}+\lambda_1\|\boldsymbol{S}-\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{T}}\|_F^2+\lambda_2\|\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T}\|_F^2+\lambda_3\sum_{k,l=1}^{N}\|\boldsymbol{w}_k-\boldsymbol{w}_l\|_2^2r_{kl} \tag{21}$$

本文用 $\tilde{\boldsymbol{S}}$, $\tilde{\boldsymbol{W}}$, $\tilde{\boldsymbol{T}}$ 来表示上一步迭代中获得的 $\boldsymbol{S}$, $\boldsymbol{W}$, $\boldsymbol{T}$ 的值, 用 $\boldsymbol{S}^*$, $\boldsymbol{W}^*$, $\boldsymbol{T}^*$来表示 $\boldsymbol{S}$, $\boldsymbol{W}$, $\boldsymbol{T}$ 当前迭代步中求得的最优解. 注意: 在当前迭代步骤中优化 $\boldsymbol{S}$, $\boldsymbol{W}$, $\boldsymbol{T}$ 时, $\tilde{\boldsymbol{S}}$, $\tilde{\boldsymbol{W}}$, $\tilde{\boldsymbol{T}}$ 视为常量. 由于:

$$\boldsymbol{S}^*=\underset{\boldsymbol{S}\in[0,1]^{N\times N},\boldsymbol{S}\mathbf{1}=1}{\arg\min}\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2s_{ij}+\lambda_1\sum_{i,j=1}^{N}\left(s_{ij}-\sum_{k,l=1}^{C}t_{ik}r_{kl}t_{jl}\right)^2 \tag{22}$$

于是:

$$F(\boldsymbol{S}^*,\tilde{\boldsymbol{W}},\tilde{\boldsymbol{T}})\leqslant F(\tilde{\boldsymbol{S}},\tilde{\boldsymbol{W}},\tilde{\boldsymbol{T}}) \tag{23}$$

然后, 由于 $\boldsymbol{W}$ 的更新采用的是精确线性搜索方法, 因此目标函数必然是非增的, 因此可以得到:

$$F(\boldsymbol{S}^*,\boldsymbol{W}^*,\tilde{\boldsymbol{T}})\leqslant F(\tilde{\boldsymbol{S}},\tilde{\boldsymbol{W}},\tilde{\boldsymbol{T}}) \tag{24}$$

最后, 我们证明 $\boldsymbol{T}$ 的更新不会增加目标函数的值.

$\boldsymbol{T}$ 的更新采用了放大上界的策略, 通过优化目标函数上界得以实现. 为简化计算, 现将前一步迭代更新得到的原函数(5)的第 2 项、第 3 项分别记为 $f_1(\tilde{\boldsymbol{T}})=\sum_{i,j=1}^{N}(s_{ij}-\sum_{k,l=1}^{C}\tilde{t}_{ik}b_{kl}\tilde{t}_{jl})^2$ 和 $f_2(\tilde{\boldsymbol{T}})=\sum_{i=1}^{N}\sum_{k=1}^{C}([\boldsymbol{XW}]_{ik}-\tilde{t}_{ik})^2$, 而当前步骤迭代更新得到原函数(5)的第 2 项、第 3 项分别记为 $f_1(\boldsymbol{T})=\sum_{i,j=1}^{N}(s_{ij}-\sum_{k,l=1}^{C}t_{ik}r_{kl}t_{jl})^2$ 和 $f_2(\boldsymbol{T})=\sum_{i=1}^{N}\sum_{k=1}^{C}([\boldsymbol{XW}]_{ik}-t_{ik})^2$. 当前迭代中, 原函数(5)放大后得到的第 2 项可记为

$$g_1(\boldsymbol{T},\tilde{\boldsymbol{T}})=\sum_{i,j=1}^{N}(s_{ij}^2+\sum_{l=1}^{C}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}t_{jl}^4/\tilde{t}_{jl}^3-4\sum_{l=1}^{C}s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}\tilde{t}_{jl}\lg t_{jl}-2s_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}+4\sum_{k=1}^{C}s_{ij}\tilde{t}_{ik}[\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{kj}\lg\tilde{t}_{ik}),$$

放大后, 第 3 项记为 $g_2(\boldsymbol{T},\tilde{\boldsymbol{T}})=\sum_{i=1}^{N}\sum_{k=1}^{C}(t_{ik}^2-2[\boldsymbol{XW}]_{ik}\tilde{t}_{ik}(1+\lg t_{ik}-\lg\tilde{t}_{ik})+[\boldsymbol{XW}]_{ik}^2)$. 取 $\boldsymbol{T}=\tilde{\boldsymbol{T}}$ 时有:

$$g_2(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})=f_2(\tilde{\boldsymbol{T}}) \tag{25}$$

又 $\sum_{l=1}^{C}[\tilde{\boldsymbol{T}}\boldsymbol{R}\tilde{\boldsymbol{T}}^{\mathrm{T}}]_{ij}[\tilde{\boldsymbol{T}}\boldsymbol{R}]_{il}t_{jl}^4/\tilde{t}_{jl}^3=(\sum_{k,l=1}^{C}\tilde{t}_{ik}b_{kl}\tilde{t}_{jl})^2$, 于是有:

$$g_1(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})=f_1(\tilde{\boldsymbol{T}}) \tag{26}$$

从而:

$$g_1(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})+g_2(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})=f_1(\tilde{\boldsymbol{T}})+f_2(\tilde{\boldsymbol{T}}) \tag{27}$$

又 $\boldsymbol{T}^*=\arg\min_{\boldsymbol{T}}g_1(\boldsymbol{T},\tilde{\boldsymbol{T}})+g_2(\boldsymbol{T},\tilde{\boldsymbol{T}})$, 从而 $g_1(\boldsymbol{T},\tilde{\boldsymbol{T}})+g_2(\boldsymbol{T},\tilde{\boldsymbol{T}})$ 的最优解不会超过其任意值的取值, 于是可以得到:

$$g_1(\boldsymbol{T}^*,\tilde{\boldsymbol{T}})+g_2(\boldsymbol{T}^*,\tilde{\boldsymbol{T}})\leqslant g_1(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})+g_2(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{T}})=f_1(\tilde{\boldsymbol{T}})+f_2(\tilde{\boldsymbol{T}}) \tag{28}$$

在当前步的迭代中, $g_1(\boldsymbol{T},\tilde{\boldsymbol{T}})$ 和 $g_2(\boldsymbol{T},\tilde{\boldsymbol{T}})$ 表示 $f_1(\boldsymbol{T})$和 $f_2(\boldsymbol{T})$放大后的函数, 从而:

$$f_1(\boldsymbol{T}^*)+f_2(\boldsymbol{T}^*)\leqslant g_1(\boldsymbol{T}^*,\tilde{\boldsymbol{T}})+g_2(\boldsymbol{T}^*,\tilde{\boldsymbol{T}}) \tag{29}$$

合并公式(28)和公式(29), 可以得到:

$$f_1(\boldsymbol{T}^*)+f_2(\boldsymbol{T}^*)\leqslant f_1(\tilde{\boldsymbol{T}})+f_2(\tilde{\boldsymbol{T}}) \tag{30}$$

即:

$$F(\boldsymbol{S}^*,\boldsymbol{W}^*,\boldsymbol{T}^*)\leqslant F(\tilde{\boldsymbol{S}},\tilde{\boldsymbol{W}},\tilde{\boldsymbol{T}}) \tag{31}$$

故随着迭代的增加, 原目标函数(5)的值是非增的, 因此目标函数是收敛的.

得证. □

3 实验分析

在这一节中, 本文对实验结果进行了详细的阐述与评估. 本文方法是在聚类思想的基础上融入了标签相关性矩阵作为先验知识, 以此完成对未知样本的多个标签的预测, 解决的是聚类问题. 此处的先验知识是指标签与标签之间的相关性强弱关系信息, 不包含显式的样本标签信息, 因此属于弱监督信息. 据我们所知, 目前还没有关于多标签聚类问题的已发表的研究工作, 因此在设计实验时, 按照最相关的原则, 我们选择了对比多标签分类模型以及多标签弱监督模型. 为了充分验证本文所提方法的性能, 本文设计了多组实验: 先考虑计算历史标签相关性矩阵 $\boldsymbol{R}$ 时, 样本数目的变化对最终结果所带来的影响; 接着, 测试模型对多标签的指派能力; 然后, 通过改变有监督方法的训练集比例, 来探索 WSMLLC 的预测能力; 最后, 对算法的收敛性和计算时间进行了比较.

3.1 数据集及实验设置

我们在公开数据集上进行实验, 包括 Emotions, Image, CAL500, VirusGO, Mirflickr, Pascal07-Gist (PG), Mirflickr-HarrisHue (MHH), Mirflickr-DenseHue (MDH)和 Mirflickr-DenseHueV3H1 (MDHV3H1)共 9 个数据集.

这些数据集覆盖了多个领域, 具体内容可在http://github.com/xiao-OY/MLC_toolbox/tree/master/dataset/matfile 和 http://lear.inrialpes.fr/people/guillaumin/data.php 上找到. 表 2列出了数据集的详细信息.

表 2 实验数据集

数据集	样本数目	特征数目	标签数目
Emotions	593	72	6
Image	2 000	294	5
CAL500	502	68	174
VirusGO	207	749	6
Mirflickr	25 000	150	24
PG	9 963	512	20
MHH	25 000	100	38
MDH	25 000	100	38
MDHV3H1	25 000	300	38

本文将数据集分为两部分: 一部分用于测试, 占比30%; 另一部分则作为监督方法的训练数据集, 或是作为历史信息为多标签学习方法提供标签相关性矩阵. 我们在每个数据集上运行 10 次, 并记录这 10 次重复实验的平均值和方差.

本文一共设计了 7 种方法与本文提出的 WSMLLC 模型进行比较.

(1) 有监督的分类方法 MLKNN[8]和 BiLAS[28]. 由于它们无法训练无标注数据, 因此只能通过调整训练样本的个数达到对比的目的;

(2) 半监督多标签预测方法 CNMF. 该方法提出采用带约束的非负矩阵分解方法学习未标注数据集与标签关系, 通过最小化输入模式和类标签的重叠之间的差异来实现类标签对无标签数据的最佳分配;

(3) 弱监督偏多标签学习方法 PAR-VLS[29]. 训练时输入特征矩阵与偏标签矩阵, 假设 1 表示样本有对应的标签, 0 表示无, 则偏标签矩阵则是在真实标签矩阵的基础上, 通过将部分 0 修改为 1 的方式引入噪声, 干扰训练过程. 如果考虑极限情况, 噪声足够大, 即将所有的 0 变为 1, 此问题就变成一个聚类问题, 且与本文待解决的问题一致. 但此偏多标签方法无法处理上述极限情况, 因此本文通过引入噪声的操作, 使其所解决的问题与本文所解决的问题趋于一致;

(4) 本文方法改编后的变体 FSLT, WSMLLC-J 和 WSMLLC-RBF, FLST 可看作是本文方法的消融实验, 与本文唯一不同的地方是: 它提前计算了样本的相似性矩阵 $\boldsymbol{S}$, 然后再进行变量的更新. WSMLLC-J 和 WSMLLC-RBF 则是更改了本文计算先验标签相关性矩阵的方法, 分别采用的是 jacard 相似度和径向基函数(RBF).

本文在 WSMLLC 和其他几个对比方法中需要设置几个重要参数. 对于 WSMLLC, WSMLLC-J 和 WSMLLC-RBF 这 3 个方法, 其参数λ_1, λ_2和λ_3主要在候选集$\{10^{-3},10^{-2},10^{-1},1,10,10^2,10^3\}$中进行调节. 类别划分的临界阈值 z 设置为 $1/C$. WSMLLC-RBF 中, 高斯 RBF 函数的内核参数σ设置为 1. 至于其他对比方法, 如果原文中已给出最优结果对应的默认参数, 则本文中采用相同默认值. 实验采用的评价标准是 F 值, 该方法以每个标签单独度量的 F 值的平均值作为多标签的度量结果, 共包含两种方式: 宏观平均(Macro$F$1)和微观平均(Micro$F$1)[30], 这两个评价标准是多标签学习中常用的指标[31]. 显然, 对于这两个指标而言, 数值越大, 效果越好.

3.2 对比实验结果

为了验证本文方法的有效性, 本文将上述 8 种多标签学习算法进行了比较, 其中, 有监督多标签分类方法的训练集比例, 除 VirusGO 外均设置为 5%(VirusGO 由于数据集较小故设置为 10%). 用于计算标签相关性的历史标签比例分别设置为{0.3,0.5,0.7}, 每个算法的对比结果在表 3 和表 4 中进行展示, ↑表示数值越大, 分类效果越好, 效果最好的结果进行了加粗处理.

表 3 Macro*F*1 在多标签预测方法上的性能比较↑

数据集	比例	MLKNN	BiLAS	PAR-VLS	CNMF	FSLS	WSMLLC-RBF	WSMLLC-J	WSMLLC
Emotions	0.3	.3025±.0712	.2999±.0634	.3489±.1080	.3086±.0588	.2458±.0431	.3455±.0503	.3629±.0168	**.3903±.0141**
	0.5				.3338±.0636	.2586±.0572	.3597±.0397	.3747±.0195	**.3982±.0134**
	0.7				.3402±.0563	.2958±.0736	.3786±.0581	.3771±.0187	**.3997±.0288**
Image	0.3	.2927±.0554	.2891±.0280	.3041±.0773	.2541±.0403	.2158±.0341	.2615±.0307	.3060±.0383	**.3314±.0191**
	0.5				.2563±.0456	.2276±.0350	.2581±.0246	.3082±.0348	**.3328±.0162**
	0.7				.2526±.0577	.2319±.0428	.2780±.0406	.3204±.0196	**.3368±.0137**
CAL500	0.3	.0821±.0115	.0695±.0096	.0517±.0123	.1491±.0036	–	.1621±.0080	.1614±.0070	**.1932±.0095**
	0.5				.1507±.0049	.1344±.0475	.1601±.0079	.1888±.0042	**.1952±.0075**
	0.7				.1520±.0038	.1467±.0079	.1592±.0071	.1626±.0035	**.1958±.0088**
VirusGO	0.3	.1453±.0759	.2109±.0604	.2193±.0764	.2121±.0365	.1679±.0870	.2020±.0436	.2580±.0358	**.2641±.0344**
	0.5				.2243±.0206	.1882±.0716	.2265±.0311	.2561±.0428	**.2709±.0306**
	0.7				.2455±.0273	.2001±.0812	.2465±.0249	.2543±.0314	**.2670±.0278**
Mirflickr	0.3	.0215±.0048	.0383±.0080	.2055±.0167	.1568±.0069	.1498±.0071	.2019±.0014	.2085 ±.0011	**.2102±.0019**
	0.5				.1564±.0079	.1527±.0050	.2022±.0018	.2087 ±.0019	**.2106±.0019**
	0.7				.1571±.0041	.1525±.0050	.2022±.0016	.2089 ±.0017	**.2097±.0022**
PG	0.3	.0408±.0131	.0262±.0061	.1127±.0171	.0987±.0101	.0722±.0144	.1075±.0021	**.1154±.0016**	.1153±.0030
	0.5				.0996±.0108	.0841±.0152	.1069±.0035	.1158±.0033	**.1159±.0023**
	0.7				.1005±.0111	.0847±.0128	.1081±.0029	.1148±.0020	**.1175±.0018**
MHH	0.3	.0352±.0104	.0402±.0065	.1711±.0110	.1384±.0014	.1383±.0019	.1505±.0018	.1701±.0014	**.1755±.0021**
	0.5				.1385±.0011	.1384±.0015	.1506±.0016	.1711±.0018	**.1757±.0013**
	0.7				.1390±.0014	.1386±.0010	.1506±.0019	.1711±.0021	**.1759±.0015**
MDH	0.3	.0356±.0040	.0425±.0065	.1519±.0234	.1012±.0106	.1061±.0017	.1580±.0020	.1652±.0045	**.1753±.0017**
	0.5				.1021±.0127	.1063±.0008	.1583±.0019	.1595±.0067	**.1756±.0014**
	0.7				.1052±.0093	.1067±.0019	.1579±.0021	.1639±.0047	**.1760±.0014**
MDHV3H1	0.3	.0428±.0066	.0532±.0059	.1664±.0091	.1022±.0023	.0995±.0024	.1472±.0019	.1703±.0075	**.1731±.0036**
	0.5				.1020±.0023	.0995±.0021	.1470±.0021	.1709±.0032	**.1728±.0035**
	0.7				.1022±.0028	.0995±.0034	.1471±.0017	**.1730±.0020**	**.1730±.0026**

表 4 Micro*F*1 在多标签预测方法上的性能比较↑

数据集	比例	MLKNN	BiLAS	PAR-VLS	CNMF	FSLS	WSMLLC-RBF	WSMLLC-J	WSMLLC
Emotions	0.3	.3778±.0557	.3882±.0490	.3886±.0971	.3180±.0618	.2533±.0448	.3522±.0503	.3677±.0172	**.3943±.0142**
	0.5				.3433±.0654	.2621±.0580	.3628±.0408	.3804±.0204	**.4031±.0136**
	0.7				.3457±.0610	.3024±.0748	.3812±.0599	.3814±.0191	**.4045±.0287**
Image	0.3	.3245±.0391	.3129±.0272	.3297±.0792	.2584±.0433	.2218±.0377	.2659±.0315	.3078±.0398	**.3350±.0173**
	0.5				.2585±.0477	.2304±.0365	.2617±.0262	.3104±.0342	**.3380±.0162**
	0.7				.2570±.0555	.2357±.0450	.2793±.0392	.3219±.0189	**.3414±.0134**
CAL500	0.3	.3182±.0182	**.3462±.0212**	.1641±.0246	.2081±.0132	–	.1805±.0121	.2072±.0084	.2633±.0135
	0.5				.2049±.0117	.1715±.0605	.1851±.0126	.2593±.0069	.2629±.0128
	0.7				.2036±.0067	.1883±.0085	.1879±.0110	.2065±.0061	.2644±.0155
VirusGO	0.3	.2611±.0893	.2209±.1007	.2901±.1198	.2772±.0407	.2029±.0884	.2038±.0447	.3041±.0363	**.3275±.0501**
	0.5				.2769±.0302	.2325±.0712	.2371±.0370	.3102±.0437	**.3385±.0447**
	0.7				.2906±.0373	.2308±.0896	.2692±.0237	.3058±.0463	**.3419±.0306**
Mirflickr	0.3	.0571±.0161	.1143±.0230	.2204±.0200	.1835±.0090	.1747±.0095	.2282±.0015	.2389±.0016	**.2457±.0026**
	0.5				.1830±.0109	.1783±.0057	.2287±.0019	.2390±.0025	**.2461±.0021**
	0.7				.1840±.0070	.1782±.0061	.2286±.0019	.2389±.0017	**.2452±.0033**
PG	0.3	**.2113±.0444**	.1848±.0220	.1238±.0243	.1124±.0124	.0751±.0177	.1217±.0023	.1315±.0022	.1322±.0031
	0.5				.1151±.0152	.0887±.0191	.1203±.0033	.1319±.0036	.1332±.0023
	0.7				.1154±.0158	.0889±.0168	.1213±.0033	.1309±.0029	.1347±.0023
MHH	0.3	.1144±.0352	.1588±.0280	.1968±.0085	.1669±.0014	.1664±.0019	.1797±.0026	.2193±.0031	**.2266±.0037**
	0.5				.1667±.0013	.1669±.0017	.1798±.0021	.2204±.0037	**.2273±.0016**
	0.7				.1673±.0015	.1668±.0010	.1803±.0019	.2207±.0038	**.2286±.0028**
MDH	0.3	.1080±.0120	.1667±.0319	.1833±.0207	.1301±.0122	.1363±.0020	.1850±.0026	.2150±.0078	**.2236±.0044**
	0.5				.1310±.0142	.1362±.0011	.1861±.0030	.2045±.0109	**.2237±.0022**
	0.7				.1352±.0091	.1367±.0018	.1862±.0026	.2132±.0080	**.2234±.0028**
MDHV3H1	0.3	.1295±.0216	.2003±.0265	.1926±.0089	.1315±.0026	.1291±.0024	.1764±.0022	.2206±.0087	**.2218±.0034**
	0.5				.1314±.0032	.1288±.0024	.1763±.0031	**.2239±.0032**	.2214±.0042
	0.7				.1316±.0034	.1290±.0038	.1765±.0020	**.2246±.0038**	.2233±0039

基于表 3、表 4 中的结果, 可以得到下述结论.

(1) 在大多数数据集上, WSMLLC 与其他多标签学习方法相比具有明显的优势, 特别是在 Emotions, Image 和 VirusGO 数据集上, 本文方法总能取得最好的性能. 在数据集 Mirflickr, PG, MHH, MDH 和 MDHV3H1 上, WSMLLC 与 WSMLLC-J 取得的结果相差不大, 主要是因为 WSMLLC 与 WSMLLC-J 采用的是相同的模型, 只是在计算标签相关性矩阵 $\boldsymbol{R}$ 时采取了不同的方法. 由于数据集 CAL500 样本数目较少, 但标签数目很多, 当计算标签相关性矩阵 $\boldsymbol{R}$ 的信息量太少时(0.3), 很可能无法得到有效结果(如 FSLT);

(2) 多标签学习有一个明显的不足: 随着标签数目的增多, 标签组合数会呈指数增长, 如 C 个标签的标签组合数有 2^C 种. 因此, 当标签数目增多时, 评价指标的预测性能将会受到一定影响, 这也是文中评价结果值偏低的一个重要原因;

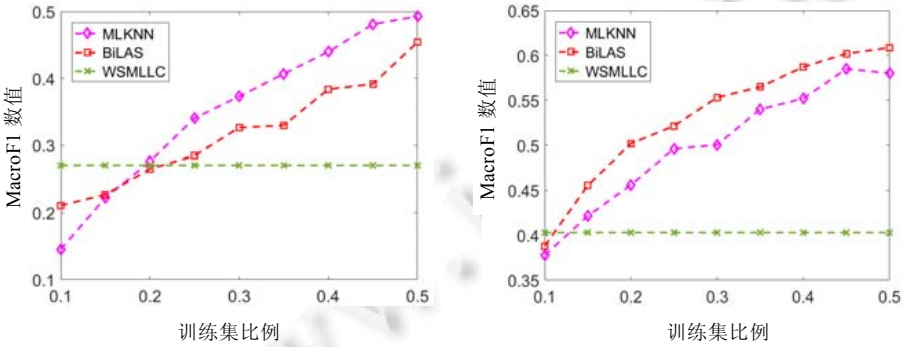
(3) 通过观察每个方法在不同数据集上的表现可以发现: 随着计算先验信息的标签数目的增加, 计算结果通常越来越好, 但这种增量较小, 这表明先验信息的计算对最后的结果虽然有一定的影响, 但这种影响相对较小, 模型比较稳定. 此外, 不同计算标签相关性的方法(即 WSMLLC, WSMLLC-J, WSMLLC-RBF)得到的结果差异有时很大, 如 Image, CAL500 和 VirusGO; 有时很小, 如 PG, 只有利用余弦相似度计算得到的结果(即 WSMLLC)是相对稳定的, 且始终处于相对较好的水平;

(4) 消融实验表明: 相比于给定样本相似度 $\boldsymbol{S}$, 边学习边更新更有助于提高模型的性能. 此外, 由于 CNMF 对先验信息和数据集的利用度有限, 尽管它的计算复杂度低, 但是最终的效果并不理想.

总而言之, 本文提出的方法在大多数情况下优于其他方法.

3.3 WSMLLC方法的标签指派能力

为了检验本文提出的多标签学习方法 WSMLLC 的标签指派能力, 本文通过改变传统有监督多标签学习方法的训练集比例, 将弱监督方法 WSMLLC 与有监督方法 MLKNN 和 BiLAS 的结果进行比较, 可以估算出 WSMLLC 标签指派能力的强弱. WSMLLC 的输入为特征矩阵 $\boldsymbol{X}$ 和标签相关性矩阵 $\boldsymbol{R}$, 其中, 计算标签相关性矩阵 $\boldsymbol{R}$ 的样本比例设为 0.5. 为了直观地看到 WSMLLC 在不同数据集上的标签指派能力, 本文将比较结果进行了可视化, 如图 2 所示.



(a) VirusGO-Macro$F1$ (b) Emotions-Micro$F1$

图 2 WSMLLC 标签指派能力可视化图

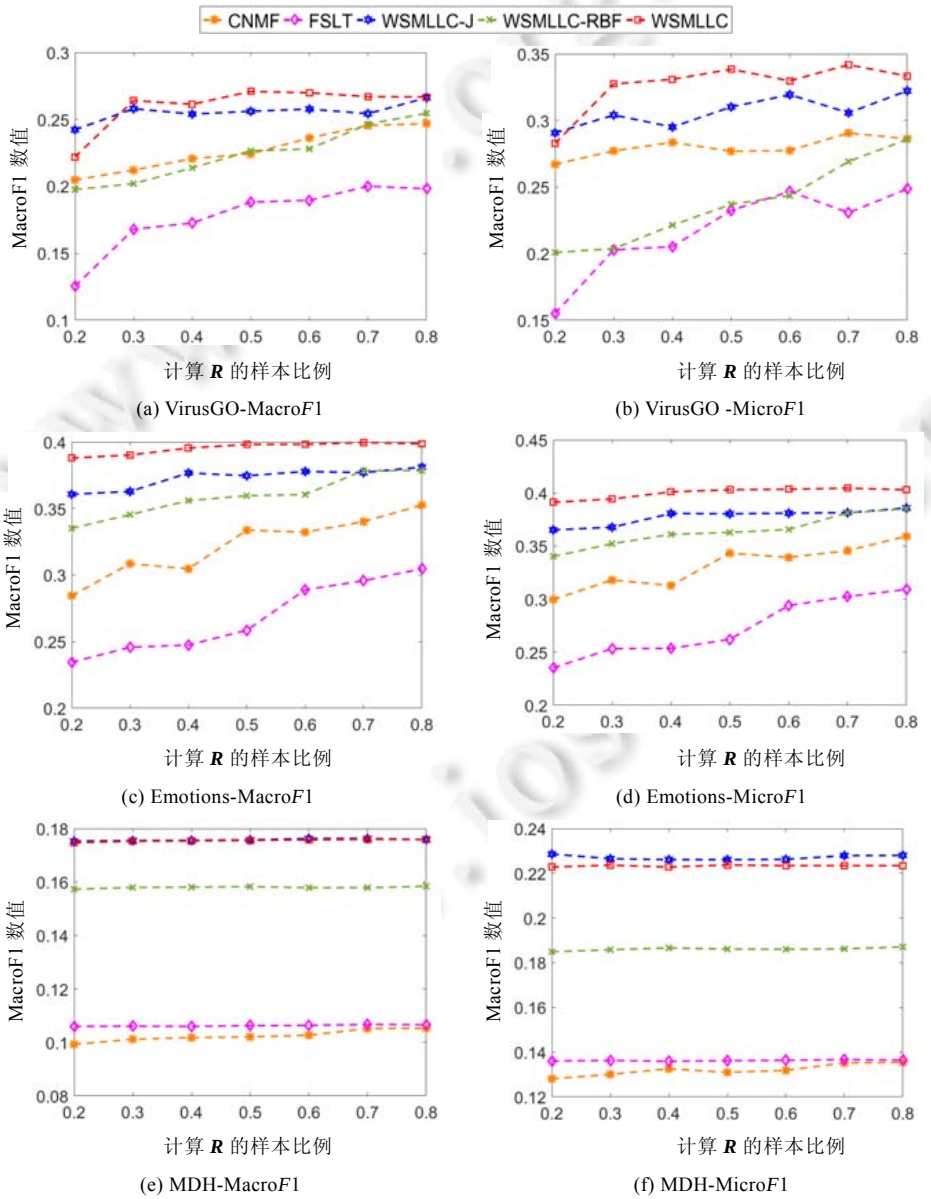
根据图 2 可以得到如下结论.

- 一方面, 随着样本在训练集中所占比例的增加, MLKNN 和 BiLAS 的结果都表现得越来越好. 这说明有监督方法受训练集影响较大, 特别是当训练样本数目发生变化时, 其相应的评价指标波动较大. 其原因是: MLKNN 模型的建立受到邻近 K 个样本的影响, 训练样本越多, 模型提供的信息就越准确, 所以效果就越好; BiLAS 的输出是基于二分类器的集成结果, 随着样本数目的增加, 每个基分类器训练的模型会更加精确, 因此结果会更好;

- 另一方面, WSMLLC 的标签指派能力在不同数据集中有所差异. 在数据集 VirusGO 中, WSMLLC 的标签指派能力相当于为 MLKNN 和 BiLAS 提供的约 20%的训练样本. 然而, WSMLLC 的影响并不总是突出的, 如在数据集 Emotions 中, WSMLLC 的聚类能力仅相当于 MLKNN 和 BiLAS 提供的约 10%~15%的训练样本.

3.4 标签比例对预测性能的影响

为了进一步探究计算标签相关性矩阵的样本比例 $\boldsymbol{R}$ 对结果的影响, 本文在图 3 中绘制了与先验信息比例 $\boldsymbol{R}$ 有关的 5 种方法随计算标签相关性矩阵的样本比例增加而变化的曲线, 其中, 计算标签相关性 $\boldsymbol{R}$ 的样本比例从 0.2 依次增加到 0.8, 每次增幅为 0.1.



(a) VirusGO-Macro$F1$ (b) VirusGO -Micro$F1$

(c) Emotions-Macro$F1$ (d) Emotions-Micro$F1$

(e) MDH-Macro$F1$ (f) MDH-Micro$F1$

图 3 标签比例改变时结果比较图

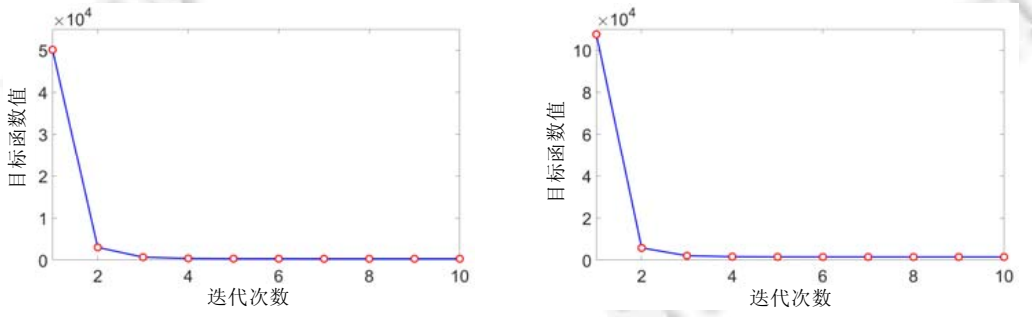
如图 3 所示, 红色标记的方法(WSMLLC)在大多数数据集中表现最好, 蓝色标记的 WSMLLC-J 效果次之, 紫色标记的 FSLT 表现基本较差. WSMLLC 和 FSLT 唯一的区别在于, 优化过程中是否持续更新了样本相关性 $\boldsymbol{S}$. 但预测结果之间的巨大差异表明, 样本相关性 $\boldsymbol{S}$ 应该随着迭代的进行而不断进行更新.

随着标签比例的增加, Macro$F1$ 和 Micro$F1$ 的值呈现出 3 种变化趋势: 一种是持续上升趋势, 如图 3(a)、图 3(b)所示, 即历史样本数目的变化对标签相关矩阵具有持续的影响, 说明此时样本数目的增加, 能够提供较多标签间的信息; 第 2 种趋势是先上升后趋于稳定, 这说明用于计算标签相关性的样本数目达到一定数量后, 相关性矩阵 $\boldsymbol{R}$ 基本保持不变; 第 3 种趋势就是不管计算 $\boldsymbol{R}$ 的样本数目怎么变化, 结果都趋于稳定, 如图 3(e)、图 3(f)所示, 这说明此时样本数目的增加已无法为我们提供更多的标签信息, 这可能与数据集样本量的大小有关, 如数据集 MDH 含有 25 000 个样本, 即使只用 20%的数据量计算 $\boldsymbol{R}$, 也有 5 000 个样本, 此时提供的标签间的信息已经十分完善, 所以此时增加样本数据量, 对结果的影响十分微小.

尽管 Macro$F1$ 和 Micro$F1$ 的值会随着计算 $\boldsymbol{R}$ 的样本比例的增加而呈上升或稳定的趋势, 但该比例的大小对最终结果的影响较小. 这意味着历史标签相关性矩阵可能比较容易获取, 且受标签数目的影响较小. 通过观察可以发现: 当计算 $\boldsymbol{R}$ 的样本比例达到 0.5 时, 最终结果基本变化不大. 因此, 后续的两个实验都默认计算 $\boldsymbol{R}$ 的样本比例为 0.5.

3.5 收敛性及时间复杂度分析

本节深入考虑了 WSMLLC 的收敛行为和计算复杂度. 如果连续两次迭代的目标函数差值小于前一次迭代的 10^{-3}, 则认为该算法是收敛的. 图 4 展示了本文方法在数据集 Emotions 和 Image 上关于 WSMLLC 的收敛曲线. 通过收敛图可以发现: WSMLLC 的目标函数值是单调下降的, 且在前两次迭代时降幅较大; 在迭代 10 次左右时, 目标函数值基本稳定.



(a) Emotions (b) Image

图 4 WSMLLC 的收敛速度

本文还进一步探讨了所提算法的时间复杂度, 并比较了算法在不同数据和特征规模的 9 个数据集上的运行成本. 实验的两个参数: 关于监督方法的训练集比例和关于计算标签相关性先验的样本的比例, 本文分别设置为 0.05 和 0.5.

表 5 总结了所有对比方法的计算复杂度, 其中, M 和 N 分别表示训练样本和测试样本的数量, D 和 C 表示数据集的维数和标签数目, η表示迭代次数, β为原文中的参数. 在所有的方法中, 可以发现, WSMLLC 的时间复杂度性能表现中等. 表 6 展示了 8 种多标签学习方法的运行时间, 可以看到: MLKNN 和 CNMF 运行速度较快, 而有监督方法 BiLAS 的运行速度则明显受到样本数目的影响. 当样本数达到 25 000 时, BiLAS 运行较为缓慢; 本文提出的 WSMLLC 运行速度中等, 既不是最快的也不是最慢的. 值得注意的是: 在数据集 VirusGO 上, WSMLLC 的运行时间远多于 CNMF. 这主要是因为 VirusGO 的特征维度 D 比较高, 且与其他数据集相比, 特征维度数远超其对应的样本个数; 而 CNMF 相对于 D 的复杂度较低; 本文方法相对 D 则是平方复杂度. 在其他数据集上, 与 CNMF 相比, 本文采用同等数量级的计算时间取得了更优的性能.

表 5 多标签学习方法的时间复杂度对比

方法	时间复杂度
MLKNN	$O(MC+M^2)$; $O(MN+NC)$
BiLAS	$O(C^2(M^2(M+D)+\mathcal{F}_B(M,[\beta M])+\mathcal{F}'_B([\beta M]))); O(C^2(D[\beta M]+\mathcal{F}'_B([\beta M])))$
PAR-VLS	$O(M^2(D+K+\eta_C)+M\cdot\mathcal{F}_{QP}(M,M)+C^2(\mathcal{F}_B(M,D))+M\cdot\mathcal{F}'_B(D);O(C^2\cdot\mathcal{F}'_B(D))$
CNMF	$O(\eta(N^2C+NC^2))$
FSLT	$O(\eta(ND^2+D^2C+DC^2+NDC))$
WSMLLC-J	$O(\eta(ND^2+D^2C+DC^2+NDC))$
WSMLLC-RBF	$O(\eta(ND^2+D^2C+DC^2+NDC))$
WSMLLC	$O(\eta(ND^2+D^2C+DC^2+NDC))$

表 6 多标签学习方法的运行时间比较

数据集	方法							
	MLKNN	BiLAS	PAR-VLS	CNMF	FSLT	WSMLLC-J	WSMLLC-RBF	WSMLLC
Emotions	0.181 4	0.725 7	0.476 2	0.649 0	1.245 9	3.220 8	1.945 1	1.146 0
Image	1.467 5	1.900 8	2.186 6	4.234 4	43.575 1	13.975 7	25.078 2	12.630 1
CAL500	3.583 0	884.034 8	13.200 1	5.292 1	5.070 2	5.183 7	4.400 3	6.519 9
VirusGO	0.261 4	22.321 4	0.329 0	0.260 0	1177.673 5	537.395 4	442.163 0	375.184 7
Mirflickr	437.508 4	4 681.683 2	192.439 8	168.415 6	221.852 8	363.281 8	344.502 7	158.346 2
PG	39.980 3	432.060 7	207.446 1	57.408 2	172.719 2	85.967 7	111.608 7	84.623 5
MHH	174.483 5	7 086.656 3	236.393 2	239.161 7	225.545 7	679.374 2	671.398 9	523.874 9
MDH	174.483 5	7 086.656 3	340.177 8	239.161 7	225.545 7	426.503 4	671.398 9	523.874 9
MDHV3H1	237.827 1	6 947.555 2	740.600 7	198.901 7	262.958 7	850.276 0	717.121 7	461.508 4

3.6 阈值z的影响分析

阈值 z 主要是在软标签矩阵 $\hat{\boldsymbol{T}}$ 硬化过程中发挥作用的. 关于阈值 z 的确定, 主要参考了二分类问题中阈值的设计思路. 二分类中, 由于只有两个类, 样本被划分到其中一个类的概率为 1/2, 因此软标签矩阵硬化时阈值 z 取值为 0.5, 大于该值时为一类, 反之为另一类. 在多标签问题中, 样本类别增加至 C 类, 独立看待每个类别时, 每个样本属于每个类别的概率为 $1/C$, 因此本文选取 $1/C$ 作为阈值 z. 为讨论 z 的不同取值对算法性能的影响, 本文在如下 6 个数据集上增加了验证分析, 实验结果表明, 阈值 z 的变化的确影响了实验性能. 在两个评价指标上, 当 z 值大于 $1/C$ 时, 实验效果明显降低; 而且当 z 值小于 $1/C$ 时, 相较 $z=1/C$ 时的结果, 算法性能均能持平或略有上升. 于是, 我们对小于 $1/C$ 的 z 值展开了详细的分析. 由实验结果可知, 通过微调 z 的取值, 本文提出方法的性能可能还有进一步提升的空间(见表 7).

表 7 阈值变化对实验结果的影响

数据集	阈值 z	Maro$F1$	Micro$F1$	数据集	阈值 z	Maro$F1$	Micro$F1$
Emotions	$2/C$	.0737±.0232	.0738±.0222	CAL500	$2/C$	.0237±.0045	.0288±.0041
	$1/C$	.3982±.0134	.4031±.0136		$1/C$	.1952±.0075	.2629±.0128
	$1/2C$	.4425±.0163	.4462±.0151		$1/2C$	.2169±.0048	.2508±0053
	$1/3C$	.4590±.0122	.4627±.0117		$1/3C$	.2220±.0043	.2537±.0048
	$1/4C$	.4644±.0089	.4687±.0096		$1/4C$	.2255±.0047	.2571±.0055
	$1/5C$	.4633±.0098	.4676±.0100		$1/5C$	.2258±.0045	.2571±.0052
Image	$2/C$	.0731±.0101	.0731±.0102	Mir Flickr	$2/C$	.0520±.0027	.0570±.0028
	$1/C$	.3328±.0162	.3380±0162		$1/C$	.2106±.0019	.2461±.0021
	$1/2C$	.3712±.0090	.3728±.0092		$1/2C$	.2325±.0014	.2577±.0017
	$1/3C$	.3801±.0082	.3815±.0084		$1/3C$	.2402±.0018	.2617±.0019
	$1/4C$	.3856±.0087	.3871±.0087		$1/4C$	.2425±.2634	.2634±.0012
	$1/5C$	.3871±.0083	.3886±.0082		$1/5C$	.2441±.0015	.2646±.0016
VirusGO	$2/C$	.0448±.0327	.0678±.0399	MHH	$2/C$	.0745±.0031	.0793±.0044
	$1/C$	.2709±.0306	.3385±.0447		$1/C$	.1757±.0013	.2273±.0016
	$1/2C$	.2906±.0194	.3187±.0245		$1/2C$	.1807±.0022	.2135±.0037
	$1/3C$	.2915±.0107	.3207±.0119		$1/3C$	.1889±.0013	.2232±.0019
	$1/4C$	.3005±.0086	.3293±.0094		$1/4C$	.1918±.0026	.2260±.0035
	$1/5C$	.2989±.0160	.3292±.0169		$1/5C$	.1804±.0021	.2220±.0030

4 总结与展望

本文提出的弱监督多标签学习方法致力于解决无标签矩阵样本的多标签指派问题, 该问题具有很强的实用意义. 本文探索了一种新的利用标签相关性先验的弱监督多标签学习方法 WSMLLC, 该方法基于相似样本具有相似标签的假设, 通过样本相似性的多维度描述和样本到标签的回归模型, 可以在仅知道历史标签相关性的前提下, 为无标记的样本分配标签. 在多个多标签数据集上的多组实验结果, 验证了本文所提出的 WSMLLC 方法的有效性. 本文通过合理利用标签相关性先验, 不仅解决了有监督分类问题中缺乏标签矩阵而无法训练模型的问题, 而且克服了无监督聚类情况下的标签对齐问题. 在今后的工作中, 我们将进一步研究因标签数量增加而导致的多标签学习效果退化与效率不佳的问题.

References:

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

[2] Elisseeff A, Weston J. A kernel method for multi-labelled classification. Advances in Neural Information Processing Systems, 2001, 14: 681−687.

[3] Xiao L, Chen BL, Huang X, *et al*. Multi-label text classification method based on label semantic information. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1079−1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

[4] Zhang JJ, Wu Q, Shen CH, *et al*. Multi-Label image classification with regional latent semantic dependencies. IEEE Trans. on Multimedia, 2018, 20(10): 2801−2813. [doi: 10.1109/TMM.2018.2812605]

[5] Wang Y, Wu YJ, Wu JZ, *et al*. Multi-dimensional tag recommender model via heterogeneous networks. Ruan Jian Xue Bao/ Journal of Software, 2017, 28(10): 2611−2624 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5132.htm [doi: 10.13328/j.cnki.jos.005132]

[6] Shan JC, Hou CP, Zhuge WZ, *et al*. Co-learning binary classifiers for LP-based multi-label classification. Cognitive Systems Research, 2019, 55: 146−152.

[7] Tsoumakas G, Vlahavas IP. Random *k*-labelsets: An ensemble method for multilabel classification. In: Proc. of the 18th European Conf. on Machine Learning, Vol.4701. Warsaw: Springer, 2007. 406−417.

[8] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7): 2038-2048.

[9] Liu Y, Jin R, Yang L. Semi-supervised multi-label learning by constrained non-negative matrix factorization. Association for the Advance of Artificial Intelligence, 2006, 6: 421−426.

[10] Huang J, Xu L, Qian K, *et al*. Multi-label learning with missing and completely unobserved labels. Data Mining and Knowledge Discovery, 2021, 35(3): 1061−1086.

[11] Luo FF, Guo WZ, Chen GL. Addressing imbalance in weakly supervised multi-label learning. IEEE Access, 2019, 7: 37463−37472. [dol: 10.1109/ACCESS.2019.2906409]

[12] Zhang ML, Li YK, Liu XY, *et al*. Binary relevance for multi-label learning: An overview. Frontiers of Computer Science, 2018, 12(2): 191−202.

[13] Weng W, Lin Y, Wu S, *et al*. Multi-label learning based on label-specific features and local pairwise label correlation. Neurocomputing, 2018, 273: 385−394.

[14] Zhao W, Kong S, Bai J, *et al*. HOT-VAE: Learning high-order label correlation for multi-label classification via attention-based variational autoencoders. Association for the Advance of Artificial Intelligence, 2021, 15016−15024.

[15] Nodet P, Lemaire V, Bondu A, *et al*. From weakly supervised learning to biquality learning, an introduction. In: Proc. of the 2021 Int'l Joint Conf. on Neural Networks. Shenzhen: IEEE, 2021. 1−10. [doi: 10.1109/IJCNN52387.2021.9533353]

[16] Frénay B, Verleysen M. Classification in the presence of label noise: A survey. IEEE Trans. on Neural Networks and Learning Systems, 2014, 25(5): 845−869. [doi: 10.1109/TNNLS.2013.2292894]

[17] Sun LJ, Lyu G, Feng SH, *et al*. Beyond missing: Weakly-supervised multi-label learning with incomplete and noisy labels. Applied Intelligence, 2021, 51(3): 1552−1564.

[18] Guo HF, Han L, Su S, *et al*. Deep multi-instance multi-label learning for image annotation. Int'l Journal of Pattern Recognition and Artificial Intelligence, 2018, 32(3): 1859001−1859005.

[19] Zeng Z, Gao N, Xue C, *et al*. Learning from audience interaction: Multi-instance multi-label topic model for video shots annotating. In: Proc. of the 24th IEEE Int'l Conf. on Computer Supported Cooperative Work in Design. Dalian: IEEE, 2021. 1075−1080. [doi: 10.1109/CSCWD49262.2021.9437805]

[20] Lv G, Xu T, Chen E, *et al*. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic Embedding. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016, 30(1): 3000−3006.

[21] Xu KX, Zhao ZY, Gu JP, *et al*. Multi-instance multi-label learning for gene mutation prediction in hepatocellular carcinoma. In: Proc. of the 42nd Annual Int'l Conf. of the IEEE Engineering in Medicine & Biology Society. 2020. 6095−6098. [doi: 10.1109/EMBC44109.2020.9175293]

[22] Huang J, Xu L, Wang J, *et al*. Discovering latent class labels for multi-label learning. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence Organization. 2020. 3058−3064. [doi: https://doi.org/10.24963/ijcai.2020/423]

[23] Cheng Z, Zeng Z. Joint label-specific features and label correlation for multi-label learning with missing label. Applied Intelligence, 2020, 50(11): 4029−4049.

[24] Zhu Y, Kwok JT, Zhou Z. Multi-label learning with global and local label correlation. IEEE Trans. on Knowledge and Data Engineering, 2017, 30(6): 1081−1094. [doi: 10.1109/TKDE.2017.2785795]

[25] Mao K, Niu J, Liu X, *et al*. Word2Cluster: A new multi-label text clustering algorithm with an adaptive clusters number. In: Proc. of the 2019 IEEE Global Communications Conf. Waikoloa: IEEE, 2019. 1−6. [doi: 10.1109/GLOBECOM38437.2019. 9013266]

[26] Boyd S, Vandenberghe L. Convex Optimization. Cambridge University Press, 2004.

[27] Fan RD, Luo TJ, Zhuge WZ, *et al*. Multi-view subspace learning via bidirectional sparsity. Pattern Recognition, 2020, 108: 107524.

[28] Zhang ML, Fang JP, Wang YB. BiLabel-specific features for multi-label classification. ACM Trans. on Knowledge Discovery from Data, 2022, 18: 1−23.

[29] Zhang ML, Fang JP. Partial multi-label learning via credible label elicitation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3587−3599. [doi: 10.1109/TPAMI.2020.2985210]

[30] Li SN, Li N, Li ZH. Multi-label data mining: A survey. Computer Science, 2013, 40(4): 14−21 (in Chinese with English abstract).

[31] Sun N, Shan J, Hou C. Multi-label active learning with error correcting output codes. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. 2019. 331−342.

附中文参考文献:

[3] 肖琳, 陈博理, 黄鑫, 等. 基于标签语义注意力的多标签文本分类. 软件学报, 2020, 31(4): 1079−1089. http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

[5] 王瑜, 武延军, 吴敬征, 等. 基于异构网络面向多标签系统的推荐模型研究. 软件学报, 2017, 28(10): 2611−2624. http://www.jos.org.cn/1000-9825/5132.htm [doi: 10.13328/j.cnki.jos.005132]

[30] 李思男, 李宁, 李战怀. 多标签数据挖掘技术: 研究综述. 计算机科学, 2013, 40(4): 14−21.

欧阳宵(1998－), 女, 硕士生, 主要研究领域为机器学习, 多标签学习.

矫媛媛(1982－), 女, 博士, 副教授, 主要研究领域为数据挖掘及应用.

陶红(1990－), 女, 博士, 主要研究领域为机器学习, 系统科学, 数据挖掘.

侯臣平(1982－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 数据挖掘, 计算机视觉.

范瑞东(1996－), 男, 博士生, CCF 学生会员, 主要研究领域为机器学习, 迁移学习.