ISSN 1000-0054 清华大学学报（自然科学版） 2024 年 第 64 卷 第 4 期 9/14
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2024, Vol. 64, No. 4 679-687

基于部分标注的自训练多标签文本分类框架

任俊飞， 朱 桐， 陈文亮

（苏州大学 计算机科学与技术学院，苏州 215006）

摘 要：多标签文本分类(multi-label text classification，MLTC)旨在从预定义的候选标签中选择一个或多个文本相关的类别，是自然语言处理(natural language processing，NLP)的一项基本任务。前人工作大多基于规范且全面的标注数据集，而这些规范数据集需要严格的质量控制，一般很难获取。在真实的标注过程中，难免会缺失标注一些相关标签，进而导致不完全标注问题。该文提出了一种基于部分标注的自训练多标签文本分类(partial labeling self-training for multi-label text classification，PST)框架，该框架利用教师模型自动地给大规模无标注数据分配标签，同时给不完全标注数据补充缺失标签，最后再利用这些数据反向更新教师模型。在合成数据集和真实数据集上的实验表明，PST 框架兼容现有的各类多标签文本分类模型，并且可以缓解不完全标注数据对模型的影响。

关键词：多标签文本分类；不完全标注；自训练

中图分类号：TP393.1 **文献标志码**：A **文章编号**：1000-0054(2024)04-0679-09

DOI：10.16511/j.cnki.qhdxxb.2024.21.006

多标签文本分类作为自然语言处理中一项基本且实用的任务，可以自动地标注与文本相关的标签，在情感分析[1]、话题识别[2]、问答[3]和网页标记[4]等许多领域都有应用。然而，由于标注体系的复杂性，标注过程中可能存在标签缺失的情况，从而形成不完全标注的数据集。这种缺失标签对模型的影响主要分为 2 方面：1）退化影响：大量缺失标签的存在导致与文本相关的正例标签数量减少，模型在少量相关标签的训练下无法学习到更加全面完整的信息；2）误导影响：大量缺失标签在模型训练过程中被当作与文本不相关的负例标签计算，从而误导模型学习到相反的信息。面向不完全标注的多标签文本分类旨在从不完全标注数据集中学习文本到相关正例标签的分类器，同时尽量缓解缺失标签对模型的影响，提升多标签分类的性能。

现有针对多标签文本分类的研究主要集中在 4 方面，分别是文本语义表示、标签间关系、标签分布和文本与标签语义链接的研究。文本语义表示的研究侧重于使用深度神经网络来提取出文本的深层语义表示[5]。标签间关系的研究通常利用标签式注意力机制[6]来建模标签相关性。标签分布的研究通过设计特殊的损失函数[7]和数据采样策略[8]来缓解数据标签分布不均衡等问题。文本与标签语义链接的研究[9]采用联合建模将标签信息融入文本。然而，这些研究都是在人工标注的数据集上监督训练，无法解决不完全标注的标签缺失问题。

本文提出了一种基于部分标注的自训练多标签文本分类(partial labeling self-training for multi-label text classification，PST)框架。首先，利用多标签文本分类模型在不完全标注数据集上训练，将训练后的分类模型称为教师模型；然后，利用教师模型给大规模无标注数据和不完全标注数据打分，以获取数据在每个标签上的得分；接着，利用双阈值策略将数据对应的标签划分为正例标签、负例标签和其他标签 3 种状态；最后，通过联合训练充分利用 3 种不同状态的标签信息对教师模型进行更新。总的来说，PST 框架可以从两方面缓解缺失标签对模型的影响：一方面，为无标注数据分配标签增加了训练模型的正例标签，进而缓解缺失标签带来的退化影响；另一方面，为人工标注数据补充标

收稿日期：2023-11-09

基金项目：国家自然科学基金重点联合项目(61936010)

作者简介：任俊飞(2000—)，男，硕士研究生。

通信作者：陈文亮，教授，E-mail：wlchen@suda.edu.cn

签减少了误导模型训练的缺失标签，进而缓解缺失标签带来的误导影响。本文分别在合成数据集和真实数据集上进行实验，实验结果表明，随着不完全标注问题加剧，多标签文本分类模型的性能急剧下降，而PST框架可以在一定程度上缓解下降速度，缺失标签越多缓解越明显；同时，PST框架对不同的教师模型都有着不同程度的改善，充分证明了PST框架的通用性。

1 预备知识

1.1 任务定义

假设 $D=\{(x_1, \boldsymbol{y}_1), (x_2, \boldsymbol{y}_2), \cdots, (x_N, \boldsymbol{y}_N)\}$表示一个人工标注的多标签文本分类数据集，其中每条数据都由文本 x_i 和人工标注的标签向量 $\boldsymbol{y}_i$ 组成。$\boldsymbol{y}_i=(y_i^1, y_i^2, \cdots, y_i^m)$，其中 y_i^k 表示 $\boldsymbol{y}_i$ 中的第 k 个标签；正例标签为1，负例标签为0，其他标签(不确定为正例还是负例的标签)为2。由于不完全标注问题的存在，$\boldsymbol{y}_i$ 可能存在缺失，因此进一步定义 $\tilde{\boldsymbol{y}}_i=(\tilde{y}_i^1, \tilde{y}_i^2, \cdots, \tilde{y}_i^m)$表示理想标签向量，其中正例标签为1，负例标签为0。$\boldsymbol{y}_i$ 与 $\tilde{\boldsymbol{y}}_i$ 的按位异或即为缺失标签向量 $\hat{\boldsymbol{y}}_i=(\hat{y}_i^1, \hat{y}_i^2, \cdots, \hat{y}_i^m)$，缺失标签为1，非缺失标签为0。

一般来说，多标签文本分类任务需要学习一个从 x_i 到 $\boldsymbol{y}_i$ 的分类器，但在真实场景中可能存在不完全标注问题，导致真实标签向量 y_i 与理想标签向量 $\tilde{\boldsymbol{y}}_i$ 不一致，从而导致 $\hat{\boldsymbol{y}}$ 的存在。而不完全标注的多标签文本分类旨在从真实的 $\boldsymbol{y}_i$ 出发学习分类器，同时要尽可能地削弱未知的 $\hat{\boldsymbol{y}}_i$ 给分类器带来的影响。

1.2 多标签文本分类

1.2.1 编码

首先将 x_i 输入预训练语言模型BERT[10]得到文本编码矩阵：

$$\boldsymbol{H}=\mathrm{BERT}(x_i). \tag{1}$$

其中 $\boldsymbol{H}=(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$，$\boldsymbol{H}\in\mathbb{R}^{n\times d}$，字符的向量表示 $\boldsymbol{h}_i\in\mathbb{R}^d$，$n$ 为预定义的最大编码字符数，d 为向量维度。

接着为进一步提取文本的语义特征，多标签文本分类模型通过设计不同的网络架构对文本编码向量进一步操作以获取文本特征向量表示 $\boldsymbol{v}\in\mathbb{R}^d$，计算如下：

$$\boldsymbol{v}=\mathrm{Net}(\boldsymbol{H}). \tag{2}$$

其中Net为提取语义特征的网络架构，可以从二维的文本编码向量提取信息，并压缩成一维的文本特征向量 $\boldsymbol{v}$。不同的多标签分类模型有着不同的网络架构，本文统一用Net表示。

1.2.2 解码

解码阶段将 $\boldsymbol{v}$ 通过标签分类层得到最终的标签表示向量 $\boldsymbol{p}_i\in\mathbb{R}^m$：

$$\boldsymbol{p}_i=\mathrm{sigmoid}(\boldsymbol{v}\boldsymbol{W}_i+b_i). \tag{3}$$

其中：sigmoid为激活函数，$\boldsymbol{W}_i\in\mathbb{R}^{d\times m}$ 为可学习权重矩阵，b_i 为偏置。$\boldsymbol{p}_i=(p_i^1, p_i^2, \cdots, p_i^m)$，其中 p_i^k 表示第 k 个标签与文本相关的概率，若大于设定的阈值，则判定该标签为与文本相关的正例标签。

1.2.3 训练

在多标签分类任务中，通常使用二元交叉熵损失函数计算损失：

$$L_{\mathrm{BCE}}=\begin{cases}-\log(p_i^k), & y_i^k=1;\\ -\log(1-p_i^k), & \text{其他.}\end{cases} \tag{4}$$

然而该交叉熵损失对每个标签的计算有着相同的权重，当标签分布不平衡时优化效果降低。本文实验对比的部分多标签分类模型通过设计不同的损失函数缓解标签分布不平衡问题，以CBLoss[11]为例，计算如下：

$$r_{\mathrm{CB}}=\frac{1-\varepsilon}{1-\varepsilon^{\mathrm{freq}}}; \tag{5}$$

$$L_{\mathrm{CB}}=\begin{cases}-r_{\mathrm{CB}}(1-p_i^k)^{\gamma}\log(p_i^k), & y_i^k=1;\\ -r_{\mathrm{CB}}(p_i^k)^{\gamma}\log(1-p_i^k), & \text{其他.}\end{cases} \tag{6}$$

其中：$\varepsilon\in[0,1)$为人为设置的超参数，fred为训练集中每个标签对应的频率，$\gamma\geqslant 0$ 为可调控的聚焦参数。

2 PST框架

2.1 框架结构

PST框架结构如图1所示。一般的多标签文本分类任务直接利用分类模型在人工标注的数据集上训练更新至模型收敛。标准的自训练框架将训练好的分类模型作为教师模型，并利用教师模型自动地对无标注数据集分配标签，最后将已分配标签的数据集与人工标注数据集混合起来训练更新教师模型。具体流程见图1中红色虚线框：1) 使用人工标注数据集来训练多标签文本分类模型，并将训练收敛的模型当作教师模型；2) 使用教师模型对无标注数据集分配标签；3) 通过预先定义的阈值将预

测的标签分为正例标签与负例标签；4）将含有正例标签的数据与人工标注数据集混合起来训练更新教师模型；5）跳转到步骤 2 执行，直到教师模型的性能不再提高或满足停止条件。

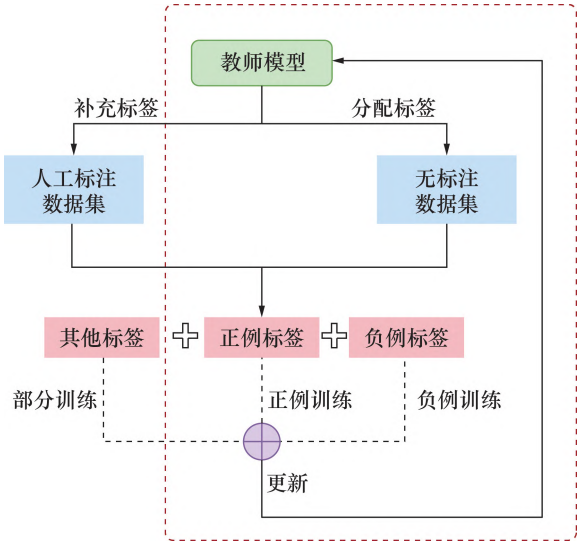


图 1 PST 框架结构图

然而在本文提到的不完全标注应用场景中，人工标注数据集可能存在缺失标签的情况，教师模型可能会受缺失标签的负面影响给无标注数据集打上错误标签。而本文提出的 PST 框架对标准自训练框架的步骤 2、3 进行了补充与修正。图 1 中，在步骤 2 利用教师模型对无标注数据集分配标签时，同时对人工标注数据集进行标签的补充预测，以补充人工标注过程中遗漏的缺失标签。为了能够同时预测这 2 部分数据集，为每条数据的每个标签定义了 1 个状态标记，防止教师模型在预测标签时破坏原有的人工标注标签。在步骤 3 通过阈值划分正例和负例标签时，利用双阈值策略将标签分为 3 类：正例标签、负例标签和其他标签。同时根据预测标签类别更新该数据对应的标签状态集合。双阈值的设定一方面能够缓解教师模型因缺失标签影响而错误预测标签的数量，另一方面通过双阈值获取的其他标签在训练时不对模型产生影响，进而缓解缺失标签导致的错误信息影响模型性能。

2.2 标签分配

本文所提出的标签分配算法见图 2，该算法利用双阈值策略和全局标签状态向量，通过教师模型为数据集补充分配标签。

具体的，本文定义了一个全局标签状态矩阵 $\boldsymbol{state}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N)$，$\boldsymbol{s}_i=(s_i^1,s_i^2,\cdots,s_i^m)$，其中 **state** 用来记录每条数据当前对应的每个标签的

输入： $\boldsymbol{Y}$ 和 D_{auto}，其中 $\boldsymbol{Y}$ 为预定义的标签集合，D_{auto} 由教师模型对人工标注数据集和无标注数据集进行打分得到。$D_{\text{auto}}=\{(x_1,\boldsymbol{p}_1),(x_2,\boldsymbol{p}_2),\cdots,(x_N,\boldsymbol{p}_N)\}$
参数： T_{Pos} 和 T_{Neg}
输出： 下一轮训练的标注数据 D_{train}

```
1  初始化 state
2  D_train = [ ]
3  for (x_i, p_i) ∈ D_auto do
4    s_i = state[i]
5    y_i = (y_i^1, y_i^2, ···, y_i^m)
6    for (p_i^k ∈ p_i and y_i^k ∈ y_i and s_i^k ∈ s_i) do
7      if s_i^k == 2 // 确定正例，无需改动
8        y_i^k = 1
9        continue
10     if s_i^k == −2 // 确定负例，无需改动
11       y_i^k = 0
12       continue
13     if p_i^k > T_Pos // 预测正例，标签状态加 1
14       s_i^k = s_i^k + 1
15       y_i^k = 1
16     elif p_i^k < T_Neg // 预测负例，标签状态减 1
17       s_i^k = s_i^k − 1
18       y_i^k = 0
19     else y_i^k = 2 // 其他标签
20   if y_i not all 0
21     (x_i, y_i) → D_train
22   return D_train
```

图 2 标签分配算法

状态，该状态用于辅助后续标签的分配，其中 s_i^k 记录当前第条数据对应的第 k 个标签的状态值，该状态值的取值空间为{−2，−1，0，1，2}。−2 代表确定负例标签状态，即确定该标签与文本不相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作负例标签；−1、0、1 代表中间状态，处于该状态的标签暂未确定是否与文本相关，PST 框架会利用教师模型对这些状态的标签预测打分，并根据得分和预定义的阈值来动态地改变标签的状态；2 代表确定正例标签状态，即确定该标签与文本相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作正例标签。首先初始化 1 个全局标签状态向量：把人工标注数据的正例标签状态初始化为 2，其他所有标签的状态初始化为 0（图 2 第 1 行）。接着，利用教师模型对所有数据进行标签预测（图 2 第 3 行）。之后，利用人为定义的正例阈值 T_{Pos} 和负例阈值 T_{Neg} 与模型预测的每个数据对应的所有标签的得分进行比较来将

标签分为正例、负例或其他3类，同时更新**state**(图2第13—19行)，其中T_{Pos}和T_{Neg}的取值根据多组取值组合实验中最佳实验结果确定。当某数据的某个标签状态为2时，即模型连续2次给予该标签高分，则该标签在后续训练过程中无需模型预测打分；若某数据的某个标签状态为-2，即教师模型连续2次给予该标签低分，则该标签无需模型预测打分(图2第7—12行)。最后通过条件判断返回用于下一轮训练更新教师模型的数据(图2第20—22行)。

2.3 联合训练

通过标签分配算法获取的新数据集将直接用于训练更新教师模型。不同于式(6)直接利用人工标注数据集训练教师模型，由于双阈值的设定，新数据集中添加了其他标签的额外信息。为此PST框架通过修改教师模型的损失函数以引入其他标签的额外信息，削弱错误信息对模型的误导。本文对式(6)进行修改得到

$$L_{\text{CB-Part}}=\begin{cases}-r_{\text{CB}}(1-p_i^k)^{\gamma}\log(p_i^k), & y_i^k=1;\\ 0, & y_i^k=2;\\ -r_{\text{CB}}(p_i^k)^{\gamma}\log(1-p_i^k), & y_i^k=0.\end{cases} \tag{7}$$

3 实验与分析

本文分别在合成数据集和真实数据集上进行实验，以证明所提出方法的有效性和通用性。

3.1 实验数据

3.1.1 合成数据集

本文采用多标签文本分类任务中常见的英文数据集AAPD[12]作为合成数据，该数据集是由网络上收集的55 840篇论文的摘要和相应学科类别组成。1篇学术论文属于1个或者多个学科，总共有54个学科类别，目的是根据给定的摘要来预测学术论文相对应的学科。为了模拟不完全标注的数据集，本文在标注规范的AAPD数据集上，按8∶1∶1的数量比例将数据集切分为训练集、验证集和测试集，并对训练集按照不同的缺失概率来随机删除一些标签。同时为了更好地评估模型的性能，未对验证集和测试集随机删除标签。

为了更全面地分析不同场景下的不完全标注问题，采用2种方案来人为删除标签构造不完全标注的合成数据集。方案1：确保删除标签后的数据集每个数据至少仍保留1个相关标签。首先统计平均每个数据包含的标签数为2.41，并进一步确定该方案下标签缺失概率最大为(2.41−1.0)÷2.41=0.585，因此分别按照p为0.1、0.2、0.3、0.4、0.5、0.585来按标签分布等比例删减相关标签并且始终保证每个数据至少有1个标签；方案2：对于标注数据集中未标注标签的数据，假设该数据被标注为与所有标签都不相关仅含负例标签的数据。按照这种假设，分别按照p为0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9随机删除标签，当较大时数据的所有标签可能都会丢失，对于此类数据，本文并未抛弃，仍当作只含负例标签的数据训练多标签文本分类模型。按照上述2种方案人为构造了15组数据集(方案1构建6组，方案2构建9组)，加上原始数据集共构建了16组合成数据集进行相关实验。

3.1.2 真实数据集

本文对CCKS2022 Task8面向金融领域的Few-Shot事件主体抽取学术评测提供的数据集进行人为修正，将其构建为多标签文本分类任务的中文数据集CCKS-IMLTC作为真实场景下不完全标注的数据集。该数据集具体的构建流程如下：1) 原CCKS数据为事件主体抽取数据集，每条数据由1段文本和该文本包含的1个或多个事件类型与事件主体对构成，但其标注质量不高，存在比较严重的事件类型缺失问题。因此本文删除原数据集中的事件主体只保留事件类型，同时删去了部分数据量极少的类别，将其修正为多标签文本分类任务的中文数据集，并按8∶1∶1的数量比例将数据集切分为训练集、验证集和测试集。2) 为了更加精确地评估模型性能，对划分后的验证集和测试集进行人工补全，并对每个补充的标签打上标记便于后续统计每个标签的缺失情况。3) 对比补充前后的测试集与验证集，估算出整体标签的平均缺失比例为9.2%。在对测试集和验证集进行缺失标签补充的过程中发现，有一部分关联性较强的标签的缺失率可达60%左右，本文将这些标签当作1个集合称为Few，并在实验中单独对这些标签进行评价。

3.2 评价指标

本文用2类指标全面评价模型性能。第1类指标为精确率P、召回率R和F1值，该类指标常用于分类任务模型性能的直观评估。第2类评价指标为退化率α_p和误导率β_p，该类指标用于评估分析缺失标注标签对模型的退化和误导2方面影响[13]，计算如下：

$$\alpha_p = \frac{f_0^{\mathrm{a}} - f_p^{\mathrm{a}}}{f_0^{\mathrm{a}}}, \tag{8}$$

$$\beta_p = \frac{f_p^{\mathrm{a}} - f_p}{f_p^{\mathrm{a}}}. \tag{9}$$

其中：f_0^{a} 表示在原始训练集上训练得到的分类模型的 F1 值，f_p 表示在按 p 生成的训练集上用式(6)训练得到的分类模型的 F1 值，f_p^{a} 表示在按 p 生成的训练集上引入其他标签信息采用式(7)训练得到的分类模型的 F1 值。在真实数据集中，由于不知道具体的缺失标注的标签信息，因此实验部分只将第 2 类指标作用到本文构建的 15 组合成数据集上。从式(8)可以发现，α_p 指标不包括缺失标签对损失的影响，因此可以评价正例标签数量减少导致分类器训练不充分的退化影响；从式(9)可以发现，β_p 指标通过计算相同缺失概率下消除缺失标签影响与未消除缺失标签影响的 F1 值差值得到，因此可以评价缺失标签对分类器的误导影响。

3.3 实验设置与对比模型

本文将 PST 框架运用到多种常见的多标签文本分类模型上进行实验，相关模型如下：CLS[10]、TextCNN[14]、LSAN[15]、FL[7]、RFL[16]、CB[11]、DB[16]、HTTN[17]、LACO[18]、FLEM[19]和本文结合 TextCNN[14]与 CB[11]构建的 TextCNN-CB 模型。

使用开源的预训练模型 bert-base-cased (https://huggingface.co/bert-base-cased)和 bert-base-chinese (https://huggingface.co/bert-base-chinese)分别作为英文和中文的编码层，依据教师模型在验证集上的性能设定超参。具体地，设置最大句长为 256，学习率为 2×10^{-5}，batch-size 为 16，教师模型的训练轮次为 20，自训练迭代轮次为 10，随机种子为 1 227，dropout 为 0.5，T_{Pos} 为 0.6，T_{Neg} 为 0.4，线性层维度大小为 300，HTTN 模型中头部标签数量为 84，TextCNN 滤波器为 200、窗口大小为[1，3，5，7]，LSTM 隐层大小为 256，其他超参均遵循相关文献的设置。

3.4 实验结果与分析

3.4.1 真实数据集上的实验

表 1 为 CCKS-IMLTC 数据集上不同多标签文本分类模型在不同框架下的实验结果，其中教师模型表示直接在原始数据集上充分训练的分类模型；自训练模型表示对教师模型采用标准自训练框架进一步更新迭代得到的模型；PST 模型表示对教师模型采用 PST 框架更新迭代得到的模型。Δ_{T} 和 Δ_{ST} 分别表示 PST 模型的 F1 值相比教师模型和自训练模型的变化。横向对比分类模型在不同框架下训练后的结果发现，标准自训练框架对某些模型的性能有所提升，然而这种提升并不稳定，而 PST 框架对所有教师模型都有较高的性能提升，充分证明了 PST 框架的有效性与通用性。对比不同框架下的性能发现，标准自训练框架通过优化模型的 P 提升性能，相反 PST 框架主要优化模型的 R 以提升整体性能。并且相比自训练框架，PST 框架对教师模型性能的提升更加稳定且有效。进一步分析实验结果发现，教师模型与自训练模型将缺失标签当作负例标签训练，因此其预测结果大都是训练中出现过的正例标签，进而导致预测出的正例标签准确率较高。而 PST 框架通过双阈值策略和全局标签状态补充缺失的正例标签，进而使得模型可以预测出更多的正例标签。

表 1 不同分类模型在 CCKS-IMLTC 数据集上的实验结果

分类模型	教师模型			自训练模型			PST 模型				
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%	Δ_{T}	Δ_{ST}
CLS	**79.91**	63.81	70.95	77.37	63.08	69.50	76.60	**67.25**	**71.62**	+0.67	+2.12
TextCNN	76.69	71.91	74.22	75.91	**73.87**	74.88	**78.06**	72.22	**75.03**	+0.81	+0.15
LSAN	75.51	59.89	66.80	**77.24**	60.89	68.10	75.34	**63.92**	**69.16**	+2.36	+1.06
FL	80.99	68.83	74.41	**82.14**	67.83	74.30	80.44	**70.38**	**75.07**	+0.66	+0.77
RFL	80.65	69.67	74.76	**81.97**	69.37	**75.15**	80.36	**70.58**	**75.15**	+0.39	+0.00
CB	81.16	70.28	75.33	**81.83**	69.39	75.10	80.51	**71.03**	**75.48**	+0.15	+0.38
DB	74.76	74.50	74.63	73.99	**76.66**	75.30	**77.32**	74.68	**75.97**	+1.34	+0.67
HTTN	81.46	67.81	74.01	**81.56**	68.38	74.39	80.79	**69.88**	**74.94**	+0.93	+0.55
LACO	78.19	**71.45**	74.65	78.85	70.46	74.42	**79.53**	70.68	**74.84**	+0.19	+0.42
FLEM	80.54	69.39	74.55	**83.91**	67.81	75.01	82.05	**69.57**	**75.30**	+0.75	+0.29
TextCNN-CB	76.56	74.35	75.44	77.00	73.53	75.22	**77.59**	**74.68**	**76.11**	+0.67	+0.89

图3为不同模型在CCKS-IMLTC中标签缺失严重的Few标签集合上的实验结果，可以发现相比表1整体标签上的实验结果，PST框架在缺失比例高的标签上对教师模型的提升更为明显。同时发现当教师模型性能过低时，标准自训练框架反而会给模型带来负优化，而PST框架对所有教师模型都有不同程度的提升，更加充分地证明了PST框架的通用性，可以兼容现有多种不同的多标签文本分类模型。进一步对比不同模型在整体标签和Few标签集合上的实验结果，发现相比注重标签间关系的模型FLEM和LACO，针对标签分布设计的模型如CB、DB等的性能随着标签缺失率升高而下降得较为缓慢。

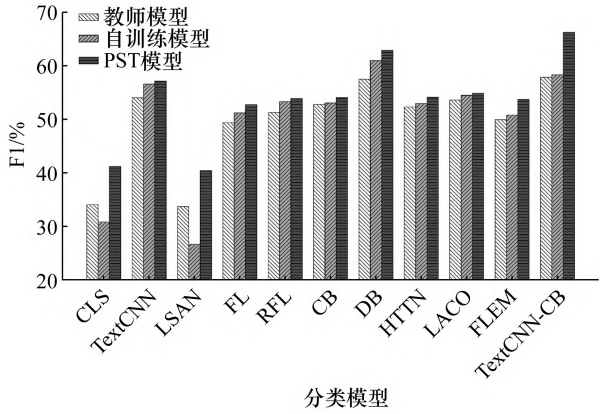


图3 不同分类模型在Few标签上的实验结果

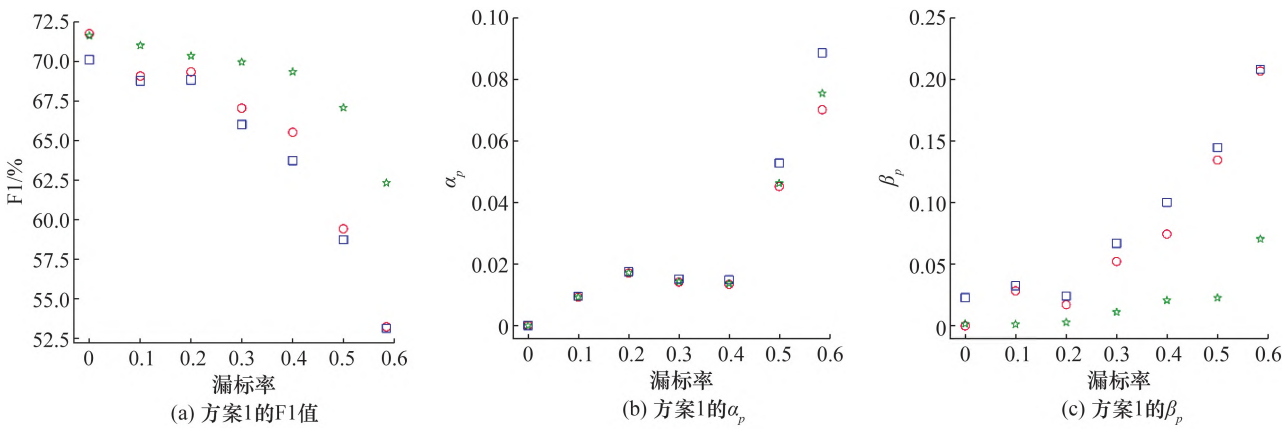
3.4.2 合成数据集上的实验

图4为TextCNN-CB在2种标签缺失方案构造的合成数据集上的实验结果。由图4a和4b的折线图，可以发现随着缺失率的增加不同框架下模型的性能都会有所下降，但相比教师模型和标准自训练框架，PST框架下降得更加缓慢，即PST框架可以有效缓解缺失标签对模型的负面影响进而提升模型性能，并且标签缺失问题越严重提升效果越明显。同时标准自训练框架在2种不同标签缺失方案下性能基本都不如教师模型，说明教师模型受到缺失标签的影响，在对训练集打分时可能会导致标注错误标签进而影响模型下一轮训练，而PST框架可以通过双阈值策略和全局标签状态来缓解这种错误标签的出现。

图4b和4e为不同框架下退化率随标签缺失比例变化的折线图，可以发现标签缺失比例不足50%时不同框架下缺失标签带来的退化影响都很小，最高仅为5%。当缺失率达到70%时，因缺失标签带来的退化影响陡增，而PST框架可以明显地缓解退化影响，尤其是缺失率为90%时PST框架可以将退化率从90%降至40%。图4c和4f为不同框架下误导率随标签缺失比例变化的折线图，可以发现随着标签缺失率的增加越来越多的缺失标签被当作负例学习误导模型，而PST框架将可能产生误导的缺失标签转为其他标签忽略其损失，进而有效地减弱模型被误导的概率。总的来说，实验结果表明PST框架可以缓解缺失标签对模型误导和退化2方面的影响。

3.4.3 不同正、负例阈值对PST框架的影响

图5为PST框架中正、负例阈值不同组合时TextCNN-CB在CCKS-IMLTC数据上的实验结果热力图。可以发现，PST框架的T_{Pos}为0.6和T_{Neg}为0.4时模型性能最优。



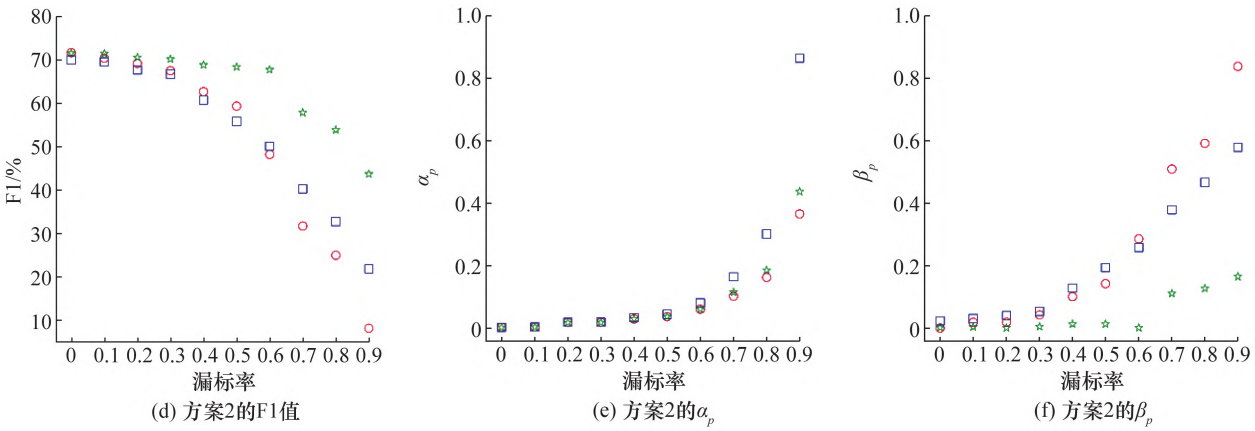


图 4 不同标签缺失方案人工合成数据集上的实验结果

3.4.4 消融实验

为了进一步分析 PST 框架中各个组件对整个框架的影响，本文利用 TextCNN-CB 分类模型在 CCKS-IMLTC 上进行了消融实验，实验结果如表 2 所示，其中 $F1_{Few}$ 表示模型在 Few 标签集合上的 F1 值，$F1_{All}$ 表示模型在全部标签集合上的 F1 值，w/o 表示去掉某一组件，数值右下角正负数表示相比教师模型性能的变化。可以发现，去掉 T_{Neg}，整体框架退化为标准自训练框架，此时 $F1_{All}$ 有所降低；去掉其他标签，此时 $F1_{Few}$ 相比教师模型提升 1.36%，但与 PST 模型相比仍有很大差距。总的来说，PST 框架可以直接在不完全标注的数据上提升教师模型性能，同时 PST 框架中的双阈值策略、其他标签设置等各组件都是整体框架中不可或缺的部分。

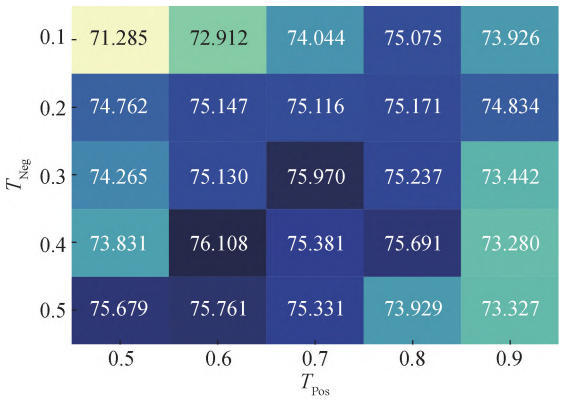


图 5 不同正负例阈值组合下的 F1/%

表 2 CCKS-IMLTC 上的消融实验 %

模型	$F1_{Few}$	$F1_{All}$
教师模型	57.90	75.44
PST 模型	$66.33_{+8.43}$	$76.11_{+0.67}$
PST 模型 w/o T_{Neg}	$58.35_{+0.45}$	$75.22_{-0.22}$
PST 模型 w/o 其他标签	$59.26_{+1.36}$	$75.37_{-0.07}$

4 结 论

本文提出了一种基于部分标注的自训练多标签文本分类框架，以缓解不完全标注问题在多标签文本分类中的影响。该框架是一种模型无关的插件式框架，可以兼容多种不同的教师模型。充分利用外部无标注数据来优化教师模型的同时，对不完全标注数据的缺失标签补充利用，进而削弱了缺失标签给模型带来的影响。实验结果表明，该框架具有通用性，并且能一定程度缓解数据不完全标注问题带来的影响。多标签文本分类模型的选择也影响着 PST 框架的性能上限，效果更好的分类模型通过 PST 框架可以更好地补充缺失标签，从而更大程度地缓解不完全标注的问题。因此如何设计更高效的分类模型是今后的研究方向。

参考文献 （References）

[1] LI X, XIE H R, RAO Y H, et al. Weighted multi-label classification model for sentiment analysis of online news[C]// Proceedings of the 2016 International Conference on Big Data and Smart Computing. Hong Kong, China: IEEE, 2016:215－222.

[2] DOUGREZ J, LIAKATA M, KOCHKINA E, et al. Learning disentangled latent topics for twitter rumour veracity classification[C]//Findings of the Association for Computational Linguistics. Bangkok, Thailand: ACL, 2021: 3902-3908.

[3] LANGTON J, SRIHASAM K, JIANG J L. Comparison of machine learning methods for multi-label classification of nursing education and licensure exam questions[C]// Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: ACL, 2020: 85-93.

[4] JAIN H, PRABHU Y, VARMA M. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016:935-944.

[5] LIU J Z, CHANG W C, WU Y X, et al. Deep learning for extreme multi-label text classification[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan: ACM, 2017: 115-124.

[6] 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[J]. 软件学报, 2020, 31(4): 1079-1089.
XIAO L, CHEN B L, HUANG X, et al. Multi-label text classification method based on label semantic information[J]. Journal of Software, 2020, 31(4): 1079-1089. (in Chinese)

[7] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2999-3007.

[8] LI Y M, LIU L M, SHI S M. Rethinking negative sampling for handling missing entity annotations[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: ACL, 2022: 7188-7197.

[9] DU C X, CHEN Z Z, FENG F L, et al. Explicit interaction model towards text classification[C]//Proceedings of the AAAI 33rd Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019: 6359-6366.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: ACL, 2019: 4171-4186.

[11] CUI Y, JIA M L, LIN T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 9260-9269.

[12] YANG P C, SUN X, LI W, et al. SGM: Sequence generation model for multi-label classification [C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA: ACL, 2018: 3915-3926.

[13] LI Y M, LIU L M SHI S M. Empirical analysis of unlabeled entity problem in named entity recognition[C]// Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: ICLR, 2021.

[14] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014: 1746-1751.

[15] XIAO L, HUANG X, CHEN B L, et al. Label-specific document representation for multi-label text classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Interna-tional Joint Conference on Natural Language Processing. Hong Kong, China: ACL, 2019: 466-475.

[16] WU T, HUANG Q Q, LIU Z W, et al. Distribution-balanced loss for multi-label classification in long-tailed datasets[C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020:162-178.

[17] XIAO L, ZHANG X L, JING L P, et al. Does head label help for long-tailed multi-label text classification[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021: 14103-14111.

[18] ZHANG X M, ZHANG Q W, YAN Z, et al. Enhancing label correlation feedback in multi-label text classification via multi-task learning[C]//Findings of the Association for Computational Linguistics. Bangkok, Thailand: ACL, 2021: 1190-1200.

[19] ZHAO X Y, AN Y X, XU N, et al. Fusion label enhancement for multi-label learning[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Shenzhen, China: IJCAI. 2022: 3773-3779.

Self-training with partial labeling for multi-label text classification

REN Junfei，ZHU Tong，CHEN Wenliang

（**School of Computer Science and Technology，Soochow University，Suzhou 215006，China**）

Abstract：[Objective] Multi-label text classification (MLTC) is a fundamental task in natural language processing. It selects the most relevant labels from the predefined label set to annotate texts. Most previous studies have been conducted on standardized and comprehensive datasets with manual annotations，which require strict quality control and are difficult to procure. In the real annotation process，some related labels are always lost，resulting in incomplete annotation. The impact of this missing label on the model is primarily divided into two forms：1) degradation effect：numerous missing labels lead to a decrease in the number of positive example labels related to the text，and the model cannot obtain more comprehensive and complete information under the training of a few related labels；2) misleading effect：numerous missing labels are treated as negative example labels that are unrelated to the text during model training，thereby misleading the model to learn the opposite information. MLTC for incomplete annotation aims to learn text from incomplete annotation datasets to classifiers for related labels while minimizing their impact on the model and improving the efficiency of multi label classification. All existing methods for MLTC involve supervised training on manually annotated data，which cannot solve missing incomplete labels. [Methods] This article proposes partial labeling self-training for the MLTC (PST) framework based on local annotation，which alleviates the negative impact of missing labels on the model by supplementing the use of missing labels. Particularly，the PST framework first utilized the basic multi label text classification model to train on incompletely labeled datasets to obtain a teacher model. Furthermore，the teacher model automatically scored large-scale unlabeled and incompletely labeled data. A dual threshold mechanism was then used to divide the labels into states based on their scores to obtain positive，negative，and other labels. Finally，the teacher model was updated using label information from three different states through joint training. To comprehensively evaluate the performance of the PST framework，we randomly deleted some labels from the training set of the English dataset AAPD，according to different missing ratios，to construct incomplete annotated synthetic datasets with different degrees of missing data. Meanwhile，we manually corrected the incomplete CCKS2022 Task 8 dataset with incomplete annotations and used it as the real dataset for the experiment. [Results] Experiments on synthetic datasets showed that as the problem of annotation intensifies，the performance of multi label text classification models decreases sharply，and the PST framework could alleviate the speed of decline to some extent，in which the more the missing labels，the more obvious the relief. The experimental results of different multi-label classification teacher models on real datasets showed that the PST framework has varying degrees of improvement on different teacher models on incompletely annotated datasets，which fully proves the universality of the PST framework. [Conclusions] The PST framework is a model-independent plug-in framework that is compatible with various teacher models. We could fully utilize the external unlabeled data to optimize the teacher model，while supplementing the use of missing labels from incomplete labeled data，thereby weakening the impact of missing labels on the model. The experimental results indicate that our proposed framework is universal and can alleviate the impact of incomplete data annotation to some extent.

Key words：multi-label text classification；incomplete labeling；self-training

（责任编辑 刘森）