层级标签语义引导的极限多标签文本分类算法

王嫄，徐涛，王世龙，周宇博，史艳翠
天津科技大学人工智能学院，天津

1 摘要

极限多标签文本分类任务具有标签集大、类间关系复杂、数据分布不平衡等特点，是具有挑战性的研究热点。现有模型对标签语义信息利用不足，性能有限。对此，该文提出一种利用层级标签语义信息引导的极限多标签文本分类模型提升策略，在训练和预测过程中给予模型层级标签引导的弱监督语义指导信息，利用这种弱监督信息规约多标签文本分类任务中要对应的多标签语义边界。在标准数据集上的实验结果表明，该文所提策略能够有效提升现有模型性能，尤其在短文本数据集中增效显著，宏精准率最高提升21.23%。

关键字: 极限多标签文本分类, 层级标签, 弱监督语义指导

2 引言

多标签文本分类将一个样本示例与一个标签子集相关联。现实世界中，数据生产速度快、体量大，具有明显的多样性和复杂性。与之对应，类标签个数也不再是以十或百为单位，而是以千、万、甚至十万为单位，典型的有知乎问答分类体系、淘宝产品分类体系等。对此，具有千个以上类标签的多标签文本分类任务，通常被称为极限多标签文本分类任务。除标签集大之外，极限多标签文本分类任务具有1)类内类间样本关系复杂，导致标签语义存在部分重叠并非完全正交，具有语义相关性;2)不同标签对应的样本数量分布高度不平衡，呈现典型的长尾分布特点。

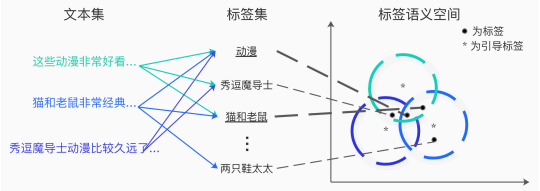

图1

本文考虑在模型训练和预测的同时，利用标签层级信息引入高级别语义范畴的引导标签。1)利用引导标签的弱监督语义信息建模标签语义相关性，隐式约束标签搜索范围;2)利用引导标签语义辐射范围下的高频标签辅助低频标签的样本学习，优化类间关系，以提升极限多标签文本分类性能。

3 相关工作

极限多标签文本分类方法早期工作重点改进传统机器学习方法以适应任务。近年，研究人员重点研究如何利用深度学习方法提升文本语义表示，如TextCNN、TextRNN、FastText和XML-CNN等，或在模型中层级标签预测，如Graph CNN和多种词嵌入组合的CNN方法等。

4 方法

在数据集中标签集内标签存在显式层级语义关系的情况下，实施引导标签策略的方法，如图2、图3所示，以基于TextCNN模型架构的策略为例。

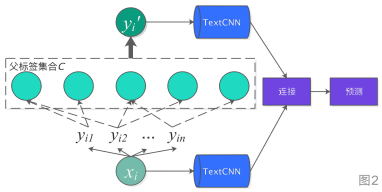

图2

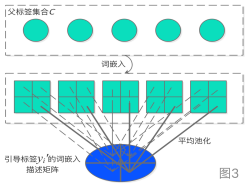

图3

对于数据集缺失标签层级信息、甚至缺失标签描述词序列的场景，本文提出启发式的伪引导标签生成策略。

具体来说，当数据集中缺失标签描述词序列时，策略通过关键词抽取方法补足描述。这种启发式的标签描述生成方法，也可结合应用采用更为复杂的描述生成方法替换。 当数据集中缺失标签层级信息时，可直接将文本对应所有标签的描述进行图3中的平均池化操作，得到伪引导标签及其对应的词嵌入描述矩阵，作为弱监督语义信息，执行双流的引导标签策略进行模型训练。

在预测阶段，会面临两种情况。第一种是待预测文本对应的层级标签集，如父标签集已知，基于该集合生成引导标签从而进行预测，常见于知乎、简书和CSDN等论坛网站中，在上传内容前，需要先选择一个大类(可对应引导标签)，内容编辑完成后，需进一步自动推荐或标记精确的标签;第二种是待预测文本对应的层级标签集未知，如新浪微博、哔哩哔哩等，本文在训练数据集合上进行近邻文本搜索，用近邻文本的引导标签作为待预测文本的引导标签。

5 实验

模型	Kanshan-Cup 数据集				Wiki10-31K 数据集			
评价指标	*macro-P*	*macro-R*	*macro-F1*	*micro*	*macro-P*	*macro-R*	*macro-F1*	*micro*
TextCNN	21.89	17.15	19.24	19.37	38.61	37.50	38.05	33.71
F-TextCNN	**31.67**	**23.81**	**27.19**	**26.36**	**41.21**	**44.59**	**42.84**	**43.51**
TextRNN	17.47	14.01	15.55	16.12	25.58	26.56	26.06	33.19
F-TextRNN	**38.70**	**24.44**	**29.96**	**27.16**	**33.82**	**28.97**	**31.21**	**34.48**
FastText	15.59	11.51	13.24	14.36	32.73	47.25	38.67	41.34
F-FastText	**28.60**	**21.49**	**24.54**	**24.45**	**35.17**	**47.39**	**40.38**	**43.86**
XML-CNN	20.37	15.19	17.40	17.93	44.30	39.25	41.62	56.88
F-XML-CNN	**29.29**	**21.19**	**24.59**	**24.27**	**45.26**	**50.26**	**47.63**	**46.27**

表2. 评价指标macro和micro性能对比
（带预测文本对应的层级标签已知）（单位：%）

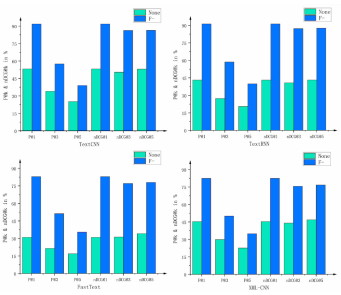

图4. 评价指标P@k和你DCG@k在Kanshan-Cup数据集上模型性能对比（带预测文本对应的层级标签已知）

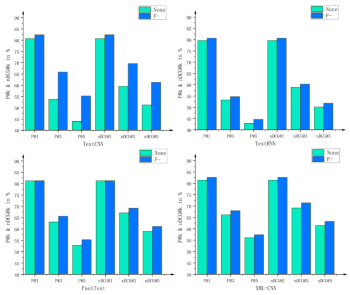

图5. 评价指标P@k和你DCG@k在Wiki19-31K数据集上模型性能对比（带预测文本对应的层级标签已知）

评价指标	*macro-P*	*macro-R*	*macro-F1*	*micro*
F-TextCNN	41.21	44.59	42.84	43.51
SF-TextCNN	40.62	46.88	43.52	43.63
F-TextRNN	33.82	28.97	31.21	34.48
SF-TextRNN	33.36	30.61	31.93	34.56
F-FastText	35.17	47.39	40.38	43.86
SF-FastText	36.64	55.67	44.20	45.12
F-XML-CNN	45.26	50.26	47.63	46.27
SF-XML-CNN	45.40	49.93	47.56	46.15

表3. 评价指标macro和micro在待预测文本对应的层级标签已知（F-）和未知（SF-）时模型性能对比（单位：%）

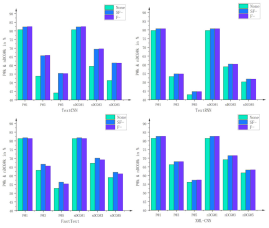

图6. 评价指标P@k和nDCG@k在Wiki10-31k数据集上模型性能对比